The Efficiency of Hospital Services and the NHS Reform:

Theory and Empirical Evidence

Alessandra Ferrari

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Economics

University of Warwick, Department of Economics

September 2001

Table of contents

List of Tables

List of Figures

Acknowledgements

I am very grateful to my supervisors Norman Ireland and Dennis Leech, whose guidance and advice have assisted considerably in the direction and completion of this thesis. I am also particularly grateful to Wiji Arulampalam for her fundamental help and advice on the empirical analysis and for accepting to be my internal advisor for the final discussion of this work.

Thanks go also to Viktor Podinovsky, Nikos Maniadakis, Bruce Hollingsworth, Pedro Pita-Barros, Prof. Marco Cicardi and Jim Storbeck for their advice on some parts of this work and for the very useful discussions.

I have also greatly benefited from the presentation of my work at the *Industrial Economics Workshop*, University of Warwick (December 1997, January 2001); the SIHCM, Third International Conference in Health Care Management University of St. Andrews (April 1998); *EARIE 98*, University of Copenhagen (August 1998); *Jornadas de Economia Industrial*, Fundacion Empresa Publica, Ministerio de la Industria, Madrid (September 1998); *Simposio de Analisis Economico*, Universitat Autonoma de Barcelona (December 1998); CMUR presentation June 2000.

Finally, I would like to thank Juan, Giuliana, Mike, Monica, Silvia and my family for listening and supporting me during these years of study at Warwick.

Acknowledgements also go to the financial support of the Teaching Assistantship scheme of the Department of Economics at Warwick in my years of study, to the Centre for Management under Regulation where I currently work and to the Statistical Office of the NHS in Scotland.

Declaration

This thesis is my own work and has not been submitted for a degree at another university.

Abstract

This thesis analyses the issue of competition for hospital services, introduced in the UK by the NHS reform in 1991.

The work is structured around two main questions: whether efficient contracts for hospital services can be devised by economic theory, and whether efficiency and productivity have actually changed since the introduction of the reform. For data reasons, the focus of this second part is on Scotland only.

Chapter 1 is a general introduction to the work.

Chapter 2 performs the theoretical analysis. The economic literature on hospital contracts in discussed first, and a model is then developed which takes into consideration the existence of waiting time and its effect on patients' utility. The conclusions cast some doubts on the possibility of defining an optimal contract, and emphasise the possible drawbacks of the prospective payment systems suggested by the reform.

Chapters 3 to 5 are devoted to the empirical analysis. As one of the main aims of the reform was to improve efficiency, this is the focus of the research, and the approach is the estimation of production frontiers, reviewed in Chapter 3.

The data are a sample of 53 acute hospitals in Scotland between 1991/92 and 1996/97 (the beginning and the end of the reform).

Two methods of estimation are used because of their complementarity: the non-parametric DEA and Malmquist indexes are the subject of Chapter 4; the econometric estimation of stochastic distance functions is in Chapter 5. The results show an improvement in productivity whereas the improvement in technical efficiency is controversial and not related to the working of the reform (represented by hospitals' trust status). Furthermore, a change in the technology of production and in what hospitals produce is found, which casts some doubts on the beneficial effects of the reform.

The general conclusions are in Chapter 6.

CHAPTER 1

INTRODUCTION

The aim of this thesis is to analyse the issue of competition for hospital services, which was introduced in the UK in 1991. The work is structured around two main questions: whether efficient contracts for hospital services can be devised by economic theory, and whether efficiency and general performance have actually changed since the introduction of the reform. For data reasons, the focus of this second part is on Scotland only.

The chapters are organised as follows.

Chapter 1 is a general introduction to the whole work.

Chapter 2 analyses the theory of contracts for hospital services, and develops a model in which waiting time is taken into consideration.

Chapters 3 to 5 are devoted to the empirical analysis: Chapter 3 reviews the literature on performance measurement; Chapter 4 carries out a non-parametric, deterministic analysis using DEA and Malmquist indexes; Chapter 5 estimates a parametric, stochastic distance function.

The comparison of the methodologies and of the results is in Chapter 6, together with the general conclusions of the work.

As regards this introductory chapter, the structure is the following: Section 1.1 is a description of the NHS reform, which motivated the research. Section 1.2 summarises some of the issues of the economic analysis of competition in health services. Section 1.3 details the aims of the work and summarises the content of each chapter.

1.1 The reform of the NHS.

Since its foundation in 1948 the NHS remained largely unchanged, in terms of organisation and objectives, until the introduction of the White Paper *Working for Patients* in December 1989. Without privatising the health service, as the sector remained public and funded by general taxation, this reform changed many key aspects of its organisation with the general aim of reshaping it in a more efficient, "business-like" way.

The reform originated in a growing sense of crisis and dissatisfaction, especially during the 70s and the 80s (Baggott, 1994; Butler, 1994). General problems, such as the existence of geographical and social inequalities in the distribution of health and health services, growing waiting lists for treatment resulting from the increasing demand, ward closures, and a general problem of underfunding, started to be attributed also to deficiencies and inefficiencies of the system itself.

In line with the government conviction about the superior efficiency characteristics of private markets, the debate about underfunding was replaced by the concern about the lack of incentives to ensure efficient resource allocation within a public system. This was seen as inflexible and ineffective in the attribution of responsibilities, fragmented among too many authorities and lacking co-ordination. Thus the apparent need to "slim" and reorganise it in a more efficient way (Holliday, 1992; Le Grand, 1994; Baggot, 1994; Bartlett and Harrison, 1993).

The White Paper *Working for Patients* was published in December 1989, and was followed by the NHS and Community Care Act in June 1990, with effect from April 1991. The general aim of the reform was to increase responsiveness to the needs of the population, to provide better services in terms of quality and quantity and a better use of the resources in terms of increased cost-consciousness.

One of the most apparent changes was the introduction of a market system (the "internal market") between providers and purchasers of hospital services, based on contractual relationships. Their roles and responsibilities would be distinguished and separated, creating a demand and a supply side: on the former, the District Health Authorities (DHAs) and the (newly created) GPs fund-holders, on the latter the new hospital trusts and private hospitals. Non trusts would continue to work as Directly Managed Units, i.e. with direct involvement of the DHA in their management. Their number was soon to fall.

Before the reform the DHAs had mixed responsibilities, covering both the planning of the services and the management of hospitals and other units. Hospitals were funded by means of a budget which basically reflected their historical costs.

After the reform, the role of the DHAs became to assess the needs of their resident population, set the priorities and ensure the availability of services, at the least possible cost and at zero cost for the patient. In their new role of purchasers of health care, they would define and conclude contracts with the providers to

buy health services. For this purpose, they would receive a budget from the DoH on a weighted capitation basis. A similar, but more restrictive, role was that of the GPs fund-holders, whose budget was top-sliced from that of the DHAs.

On the providers' side, the main change was the creation of the self-governing hospital trusts (see also Bartlett and Le Grand, 1994). Although still public and accountable to the Secretary of State, they were set as independent institutions and were given a greater autonomy in the management of their resources. Within limits[1], they could freely use and dispose of their capital assets, set wages and decide the level and composition of their staff, retain surpluses and build up reserves with which to improve their services and finance investments. Their funding would now come from the contracts concluded with the purchasers. The idea behind the reform was that competition among hospitals in order to get contracts would lead to efficiency gains.

In the words of the *White Paper*:

"...The Government believes that self government for hospitals will encourage a stronger sense of local ownership and pride...It will stimulate the commitment and harness the skills of those who are directly responsible for providing services. Supported by a funding system in which successful hospitals can flourish, it will encourage local initiative and local competition. All this will in turn ensure a better deal for the public, improving the choice and quality of the

[1] Different limits have been imposed to the autonomy of trusts, by the White Paper and after that, as regards their borrowing facilities, their pricing systems and costing procedures (prices must be based on average costs) etc.

services offered and the efficiency with which those services are delivered..." (*White Paper,* 1989, point 3.3).

As regards the kind of contracts to be used, the law was however extremely general. Three possible types were mentioned. Simple block contracts were suggested for the very beginning, a period that the Government referred to as "steady state": the meaning was that the main changes were not expected to take place all at once, in order to give time to the system to get used to them more gradually. A block contract consisted of the payment of an annual fee to the hospital in exchange for a defined range of services and facilities. This required very little information and detail and recreated a situation very similar to the one preceding the reform, with the major difference though of leaving all the risk on the provider (a problem that translated for example into the denial of access to hospital care because of the exhaustion of all available funds before the end of the year).

The other two kinds of contracts were cost-per-case and cost-and-volume contracts. With the former, the hospitals receive a fixed price per case treated, with the latter this price varies with the volume of work provided. For both, the cases are all classified into different diagnostic categories (like the American Diagnostic Related Groups, or DRGs), and the informational requirements are more complex. The only guidance given by the law in this respect was a general (and almost meaningless) reference of the price to be set equal to the average cost. These last two kinds of contract became increasingly common (Appleby, 1994). Not surprisingly, the analysis of the different incentive properties of

different kinds of contracts became of great interest. This will be discussed in Chapter 2.

The victory of the Labour Party at the general elections in 1997 and in 2001 brought some changes to the system again, substituting "competition" with "co-operation". However, the separation between health authorities and hospitals would be maintained, as well as many of the features of the internal market.

The political debate on the optimal organisation of (health and) hospital services is therefore still ongoing, and it is an issue not only in the UK.

1.2 Some economic issues.

The link between competition and efficiency is not so obvious in the case of health services. The literature on the economics of health care points to some characteristics of health provision that lead to market failures (see for example Culyer, 1971, 1991; McGuire *et al.* 1988). Together with equity considerations, this is one of the reasons why traditionally in many countries health care is publicly funded and provided (Hoffmeyer and McCarthy, 1994).

The main points can be summarised as follows.

First of all, health care is typically characterised by a high degree of uncertainty about future needs and future events[2]. The argument of bounded rationality on the conclusion of contracts is especially true for this sector: illnesses and their seriousness, or the kind of services that will be needed cannot be exactly foreseen and specified in a contract. This can translate into costly renegotiations or disputes, i.e. into high transaction costs. As shown by the early experience of simple block contracts in the UK, uncertainty led to the refusal of treatment for exhaustion of funds. The problem is made worse the higher is the degree of risk aversion of the agents.

Informational requirements are the source of other market failures. The implicit requirement of perfect information of a perfectly competitive market is in fact particularly hard to meet in the case of hospital services.

Providers must be able to exactly define and cost their activities in order to price them correctly. Cost allocation is particularly complicated because of the multiproduct nature of the service and the intrinsic problems in the definition of output itself (Elwood, 1996). This can increase the administrative costs of the service in general, a feature that became apparent quite soon after the reform as UK hospitals were not used to such detailed pricing and had to invest in specialised staff and information technology.

Purchasers must be able to observe and compare the kind and quality of the services they are offered in order to make efficient choices. Hospital services (health services in general) can be instead characterised by a high degree of

[2] For a general discussion of this topic, see for example Chalkley and Malcomson (1996);

asymmetric information between supplier and demander, be this the final consumer (the patient) or its agent on the quasi-market. The multiplicity of services offered and the various dimensions to quality are very difficult and/or very costly to observe and monitor.

The existence of asymmetric information to the advantage of the provider can create space for his opportunistic behaviour and moral hazard: for example in terms of supply-induced demand, especially in a pure market system, and/or of inefficient service-mix and choice of quality level. The first is a widely discussed problem in the US literature because of their private, insurance-based system; the second is at the centre of the debate on optimal contracting in general, as will be seen in Chapter 2.

Other issues question the possibility of competition (Propper, 1994, 1996). The hospital sector is characterised by high sunk costs, both in terms of general capital requirements and asset specificity. The existence of possible economies of scale and/or scope would make it even more unlikely (as well as inefficient) to have many providers.

The literature on this point as regards the UK is not conclusive (Bartlett and LeGrand, 1992), and in general the literature on economies of scope and scale in the hospital sector shows different and contradictory results (Butler, 1995). However, evidence from the US (Propper, 1994) suggests that a tendency towards oligopolistic or even monopolistic provision can be the result of other factors, like the purchaser's concern for quality and its unobservability: it is

---

Propper (1994); Smith and Wright (1994).

preferable to renew the contracts to the incumbents, in order to avoid the high search costs and the risk associated to a change of provider.

For the UK, this is expressed by the concern that DHAs, instead of shopping around, might just continue to send their patients to the local provider with whom they have been working for years.

It appears from the above that the benefits of competition in this sector are open to debate under many respects. The next section will clarify what issues in particular are considered in this thesis.

1.3 The aim of the thesis.

The summary of economic issues presented in Section 1.2 makes it clear how big a debate surrounds the introduction of competition in health markets.

The aim of this thesis is to focus in particular on the issue of the efficiency of competition for hospital services, both from a theoretical and an empirical point of view: does a contract exist in theory that creates an efficient market? Has the market actually improved its efficiency and productivity since the introduction of the reform?

The theoretical analysis is developed in Chapter 2. Given that a contractual system was introduced by the reform, the aim of the chapter is to discuss if a contract can be devised that gives hospitals the correct incentives so as to maximise social welfare.

An overview of the literature on the contracts for hospital services is given first. This identifies the central problem in the trade off between incentives to cost reduction and incentives to quality. In terms of payment systems, this translates into the conflict between prospective payment systems, that pay the hospital a fixed price per unit of output (price-per-case contracts), and cost reimbursement rules. A general, basic model is presented first, that defines the main set up of the problem and its possible solutions. Next, other models are discussed that relax different assumptions: observability of costs, heterogeneity among patients and/or providers, observability of quality, its possible multiple dimensions etc.

Different papers make different assumptions and describe different situations. Overall however they point to the fact that a simple price-per-case system might not lead to the identified first best solution.

After the review, a model is developed that introduces the issue of waiting lists into the basic framework. The existence of waiting time is taken into consideration by the literature mostly to explain the demand for private insurance. The chapter instead models it as a measure of demand, taking into consideration the effect that it has on patients utility and therefore on social welfare.

The general results show that a price per person demanding treatment is to be preferred to a price per case. The payment of a price per case only gives incentives to treat patients, but leads to the choice of too low a quality level. The

logic of the result lies in the mechanism that leads to the equilibrium: rewarding hospitals for their demand gives them direct incentives to increase the quality level of their service but also incentives to treat the right number of patients, because a too long waiting time in turn reduces demand. The empirical analysis shows an interesting result in this respect: a pattern over time for hospitals towards treating people on a day basis or directly as outpatients. No quality measure is available for the empirical analysis, but the concern that this phenomenon might also imply a lower quality than optimal is not uncommon.

A few extensions are made to the waiting lists model, in particular the hypothesis that social and private costs are different. When the so-called "social cost of the loss" is introduced, the model shows that a separate reward also for the number of patients treated is necessary.

The basic assumptions of the model about cost and quality observability are not relaxed in the chapter, and are left for future research.

The complexity of the identification of an optimal contract shown by the theory leads to the second question discussed in the thesis. Given that contractual agreements were used, what properties they showed in reality is a relevant point. The empirical analysis is more limited in its scope in the sense that not all the issues raised in theory can be tested in practice (this is especially true for the ones regarding quality). However, one of the main aims of the reform was to eliminate any waste of resources; the focus of the empirical analysis is therefore

on the changes in the efficiency and productivity of hospitals. As data were available only for Scotland, the analysis is restricted to that country.

More in particular, the focus is on technical efficiency, as opposed to general cost efficiency. The latter is not considered partly because others have already done it (see references in Chapter 3, Section 3.5), and partly because the definition of the price variables is particularly difficult in the case of hospital services. This would have made the (possible) analysis of the technical efficiency component more difficult and unreliable.

Focussing on technical efficiency only allows us to see whether the claimed waste in resources has actually been reduced. Insights are possible about the technology characteristics of the sector, what hospitals produced and how they produced it.

The empirical analysis is developed in Chapters 3 to 5, and it consists of the estimation of production frontiers. The data are a panel of 53 acute hospitals in Scotland between 1991/92 and 1996/97, and were obtained from the Statistics Division of the NHS in Scotland.

Chapter 3 is a review of the literature on the estimation of production frontiers. This literature is divided into two main streams, the non-parametric, mainly deterministic one, and the parametric, mainly stochastic one. The chapter analyses them both, presenting the various model definitions and characteristics. Applications of frontier models to the hospital sector are analysed at the end of

Chapter 3, with particular attention to the applications to data sets from the UK. The main differences and contributions of this work are detailed then and will be summarised shortly.

The two approaches described in Chapter 3 show almost opposite and complementary characteristics, which are discussed more in detail in Chapter 6. For this reason it is considered that a thorough analysis should make use of both of them in order to make a picture as complete and reliable as possible.

The non-parametric approach is the subject of Chapter 4. Data Envelopment Analysis (DEA) is used to estimate production frontiers and to calculate Malmquist indexes of total factor productivity. In contrast, Chapter 5 uses the econometric, stochastic approach. In particular, because of the multiple output nature of the service, a stochastic distance function is estimated. This is a relatively recent approach, not previously applied to the hospital sector. The results from the two different analyses are compared in Chapter 6 in order to draw some general conclusions.

The main contributions and differences of the empirical analysis of this thesis can be summarised as follows. The data set used covers the whole period between the introduction of the reform and the new changes introduced after the victory of the Labour party in 1997. Apart from different decisions regarding the observations to use, which might have incorrectly led previous papers to more optimistic conclusions, both the parametric and non-parametric approach are used. This allows us not only to double-check the reliability of the results, but

also to deepen the analysis itself. The specific changes in performance measured by DEA can be analysed further and related more strictly to changes in the technology of production and more generally to different choices as of what hospitals produce.

Thus, this thesis tries to address the question of the efficiency implications of the reform from both a theoretical and an empirical perspective. The overall conclusions of the work, regarding both theory and empirical analysis, are included in the final Chapter 6.

CHAPTER 2

CONTRACTS FOR HOSPITAL SERVICES

The aim of this chapter is to discuss the first question of the thesis: can a contract be devised for hospitals that leads to a socially optimal outcome. A model will be developed from the framework of Chalkley and Malcomson (1995a). This model is intended to include the existence of waiting time, as a measure of patients' demand and considering that waiting time affects patients' utility.

The chapter is structured as follows.

Section 2.1 provides an overview of the literature on hospital contracts.

Section 2.2 describes the C&M model which is the basis for the waiting lists model developed in Section 2.3. Some further extensions to the model are discussed in Section 2.4 and general conclusions are in Section 2.5.

2.1 The economic analysis of contracts for hospital services.

The key change introduced by the reform was the creation of the internal market, regulated by a system of contractual relationships between hospitals and the purchasers of their services. As mentioned in Chapter 1, the law was extremely vague about which kind of contracts to use. At the beginning, simple block contracts would have been easier because of the low informational requirements. However the idea was to move towards prospective payment systems that would pay the hospital a fixed price per case treated. This involved the definition of proper categories for the various kinds of cases treated, similar to the American DRGs. This apparently simple requirement was actually quite complicated: the

definition of cases is far from immediate; moreover, similar cases can be treated differently by different hospitals and therefore have different average costs, so that an agreed categorisation would not be that straightforward.
Nevertheless, a contractual system had to be introduced and, given the generality of the law, considerable interest developed about the different incentive properties of different kinds of contracts.

There is a very large literature[1] on the incentive properties of different payment systems to hospitals. Most comes from the USA, especially after the introduction in 1983 of a prospective payment system, which replaced the traditional cost reimbursement rule. A good review of the main issues of contracting in the NHS is in Barker et al (1996) and in Chalkley and Malcomson (1995c, 1996).

The characteristic which is common to all this literature is that of dealing with a multitask agency problem (Holmstrom and Milgrom, 1991) which is expressed in at least two potentially conflicting objectives: reducing the costs of the service without reducing the quality level.

Cost reimbursement rules, that pay the hospitals on the basis of the costs they actually incurred, lack the incentives for cost reduction. The main alternative are prospective payment systems, that consist of the payment of a fixed price per unit of output. The hospital is made the residual claimant for its costs and this gives it powerful incentives to reduce them (Laffont and Tirole, 1993). However, this can

[1] A very good survey of the issues discussed by the literature is in Chalkley and Malcolmson (1998)

lead to too low a quality level, as quality is costly but usually unmeasurable and unenforceable by contract.

This is the key problem discussed by the literature. The solutions to it differ only marginally and they depend on several things: the characteristics of the system considered (for example, if demand comes from privately insured patients), the assumptions made on the degree of asymmetric information between purchasers and providers (for example about their costs and/or about the quality level of the service), the heterogeneity of providers or patients, and so on.

A good general set up of the problem is given by Chalkley and Malcomson (1995a). The model describes the contractual relationship between a self-interested provider, that maximises their financial surplus, and a purchaser that maximises a social welfare function which depends on the number of cases treated and on the quality level of the treatment. The original demand for treatment comes from an exogenously given number of patients but the actual, final demand comes from the purchaser (in the UK this would correspond to the Health Authority or the GP fundholder) who buys hospital services for them.

Costs are assumed to be known to the purchaser. Most importantly, it is assumed that demand correctly perceives the quality level of the service. As demand is filtered, or represented, by the purchaser this is equivalent to assuming that it is this latter to have such capability.

In this framework, a prospective payment that rewards hospitals on a price per case basis is not optimal, unless the system is demand constrained, because it leads to the treatment of too many patients; the first best solution is achieved by rewarding hospitals separately for the people they treat and for those that ask for treatment. The result holds true under different capacity constraints as well as when assuming different quality dimensions.

The importance of the perception of quality by demand is not new to the literature. Its possible perverse effects have been often emphasised for the case of the US market. In the US system, where demand is fully insured, hospitals started competing on the quality of services that mattered to patients in order to attract them. Their misperception of the real quality level of the service translated into a waste of resources in amenities and facilities or unnecessary treatments, which in turn raised average costs.

The hypothesis that buyers might not infer quality correctly is also analysed in Chalkley and Malcomson (1995b). All other assumptions remaining the same as in their previous paper, no contract solves the trade off between low costs and low quality. Only a second best solution can be achieved, by means of a mixed system that combines the fixed price per case to a partial cost reimbursement.

The possibility that hospitals might not be maximising their financial surplus, but have ethical and altruistic concerns, as suggested by Newhouse (1970), is next explored in their (1995b) paper. Simple block contracts can now achieve welfare maximisation, but only if demand can be correctly predicted by the purchaser

(which is the equivalent of always having a waiting list at the end of the time period, so that hospitals always work at full capacity), or if hospitals can finance their debts. If this is not the case, then hospitals might decide to treat too few people, and cost and volume contracts are to be preferred.

Similar considerations about the "benevolence" of hospitals and how they could make sub-optimal decisions from a social welfare point of view are also in Rogerson (1994). Patients' heterogeneity is taken into consideration, and the proposed solution is a prospective payment system that price differentiates according to different demand elasticities. This is more suitable to the insurance-based US market, which is in fact the focus of the analysis.

Ma (1994), using a model in which no explicit role is given to the number of patients treated so that everything is expressed in terms of quality levels, shows that a prospective payment system is efficient also when considering the distribution of patients' health conditions. A mixed system of fixed price and partial cost reimbursement is suggested when allowing for the possibility of dumping, that is if hospitals can refuse treatment to the most expensive patients.

Similar conclusions are reached by Ellis and McGuire (1986, 1991) and by Ellis (1998): a mixed system including some cost reimbursement. Rather than talking of quality levels Ellis and McGuire focus on the multiplicity of services offered by the hospital once the patient has been admitted (like Rogerson, 1994). Ellis (1998) extends the analysis into a Cournot-Nash model in which hospitals can choose to cream skim, skimp or dump patients. Again, a mixed payment system

is proved superior to simple prospective payment and cost reimbursement. Focussing on the types of services provided rather than on quality is equivalent for the rationale of the problem. What is more relevant is that again these models are built to explain the conditions of the US market, where demand comes from fully insured patients potentially asking for an excessive level of health services. Incentives have to be considered also for the demand side, and the mixed system is the result of this interaction.

Quality discrimination among patients is the focus of the analysis of Allen and Gertler (1991) in a paper that supports the process of horizontal integration between demand and supply side, as in the case of the American HMOs (Health Maintenance Organisations). Again, a reality that is very different from the UK case.

The existence of asymmetric information about the costs of the provider is emphasised by De Fraja (2000). A regulatory model is developed, that allows for heterogeneity among providers, among patients and for the possibility of dumping. The results show that a simple price per case is not optimal and the solution is to devise a contract that links higher prices to a higher throughput. The logic of the result is that a higher price will give an incentive to more efficient hospitals to treat more expensive patients. If the price is set correctly this incentive will not work for less efficient hospitals, as the increase in revenue will not cover them for the increase in costs.

The above review is clearly not exhaustive of the whole literature, but is rather an overview of what the main issues are and how they are usually dealt with. The superiority of prospective payment systems to cost reimbursement ones is acknowledged in all cases, and some empirical evidence of their cost saving efficiency in the UK has been provided (see for example Propper, 1996). However, depending on the particular assumptions made, more complex forms of payment are suggested. No theory has yet been developed that was able to tackle all the different issues at the same time.

An issue that has not been given much consideration in the analysis of contracts is that of waiting time[2]. In the contracting context, the existence of waiting lists has been usually considered as an indirect measure of quality, often to explain the demand for private insurance (Besley *et al.*, 1999; Propper 2000).

This will be the topic of the next section. The model of Chalkley and Malcomson (1995a) (C&M from now on) will be extended to analyse a situation in which an explicit role is given to the existence of waiting time; this will be considered as a measure of hospital demand, taking into consideration the adverse effects that it has on patients utility, and therefore on social welfare.

[2] A more "technical" health economics literature on the issue exists (see for example Propper, 1995), but the perspective and approach are very different from those of this thesis.

2.2 The C&M model.

The situation modelled by C&M is the definition of a contract between a purchaser of hospital services and a self-interested hospital. The purchaser buys the services for the patients, and it maximises a social welfare function. The hospital maximises a utility function that is increasing in its financial surplus[3]. The contracting system and the determination of the purchaser's budget are taken as given, and no ethical or professional considerations enter the hospital's utility function.

The logic and the structure of the model can be summarised like this. The two objective functions are identified (that is, the purchaser's social welfare function and the hospital's utility function) and maximised with respect to the relevant variables, so that two sets of first order conditions (F.O.C.) are identified. As the hospital's utility function depends on the (kind of) payment it receives, the point is to choose a payment system that equates the two sets of F.O.C., so that the hospital makes a choice that maximises social welfare. This result is defined as a first best solution.

More in detail, define $b(x,q)$ as the benefit perceived by the purchaser of treating x patients with quality level q. This is increasing in both variables and strictly concave in q.

$C(x,q,f)$ is the cost for the hospital and is assumed to be increasing, convex in x and strictly convex in q; f is the effort in cost reduction. The cost is assumed to

[3] For a discussion of models of hospital behaviour see for example Newhouse (1970), Pauly and Redish (1973), Butler (1995) or C&M (1995b).

be known to the purchaser, whereas q cannot be monitored and so cannot be enforced by contract.

Utility for the hospital is given by

$U = s - v(x,q,f)$

where s is the financial surplus and v a disutility component. Reservation utility for the hospital is $\bar{v}$.

The social welfare function can be written as

$$W = [b(x,q) - s - C(x,q,f)] + [s - v(x,q,f))] - \alpha[C(x,q,f) + s] \quad (2.1)$$

where $\alpha > 0$ is to allow for distortions from raising revenue from taxation. The purchaser is maximising (2.1) subject to the constraint

$$s - v(x,q,f) \geq \bar{v} \quad (2.2)$$

Rearranging (2.1) and observing that it is a decreasing function of s, so that (2.2) will always hold as an equality, it can be rewritten as

$$W = b(x,q) - (1+\alpha)[C(x,q,f) + v(x,q,f)] - \alpha\bar{v} \quad (2.3)$$

The hospital objective function is to maximise

$$H = s - v(x,q,f) \quad s.t. \quad B(x,y,C) - C(x,q,f) - s = 0 \quad (2.4)$$

$B(.)$ is the total payment received by the hospital and it can be made dependent on the number of cases x, on the demand $y(q)$, which is assumed to be increasing in quality, or on costs depending on the contract chosen[4], but it cannot be directly a function of q. Thus hospital revenues can be affected by quality only via the effect that this has on demand.

The model consists of maximising (2.3) and (2.4) with respect to x, q and f under the further constraint that

$$x \leq \bar{x}(q) \qquad (2.5)$$

where $\bar{x}(q)$ is the maximum number of patients that can be treated by a hospital and is given by

$$\bar{x}(q) = min[y(q), k]$$

that is the minimum value of demand $y(q)$ and the capacity (maximum possible supply) of the hospital, k.

If (2.3) subject to (2.5) is maximised at x^*, q^* and f^*, the problem is to find the payment system $B(.)$ that maximises (2.4) subject to (2.5) at exactly the same values. Mathematically, this is done by substituting the solutions to the purchaser's F.O.C. in the hospitals' F.O.C. A first best solution exists if there is a payment system that guarantees that the equations hold. This first best solution means that hospitals will choose to treat the optimal number of patients, at the optimal quality level.

The following results are obtained by C&M.

1) For the case of a demand constrained system, that is when

$$x^* = y(q^*) < k$$

the payment of a lump-sum transfer T[5] and the use of a price per case contract of the form $B = px$ leads to the desired result if the price $p = \partial B/\partial x = B_x$[6] is set as

---

[4] For example with a cost per case contract $B = px$; with cost reimbursement $B=C(x,q,f)$, with partial cost sharing $B = px + qC$, and so on.

[5] This is required to ensure that equation (2.2) always holds as an equality.

[6] All variables with a subscript represent a derivative

$$B_x = b_x - \alpha(C_x + v_x) + \frac{b_q - \alpha(C_q + v_q)}{y_q} \qquad (2.6)$$

evaluated at the socially optimal values. That is, (2.6) reflects the marginal benefit of treating an additional patient and the marginal benefit of quality, calculated from the marginal increase in demand, all net of the cost of tax distortion.

2) If the system is unconstrained or capacity constrained, that is respectively

$x^* \leq \bar{x}(q^*)$

and

$x^* = k < y(q^*)$

then the optimal pricing rule will be given by

$$B_x = b_x - \alpha(C_x + v_x)$$
$$B_y = \frac{b_q - \alpha(C_q + v_q)}{y_q} \qquad (2.7)$$

Two different prices are in (2.7), a price per case B_x and a price per patient demanding treatment B_y. The optimal price per case equals the marginal benefit at the number of cases treated; the optimal price per patient demanding treatment reflects the marginal benefit of quality for the marginal patient; both prices are net of the marginal cost of tax distortion.

The rationale of the results is the following. A fixed price per case set as in (2.6) leads to the efficient outcome. Hospitals increase the quality level up to q^* in order to attract patients, and treat all the patients demanding treatment at that quality level. This works only if the system is demand constrained. In the case of

an unconstrained system, the optimal number of cases to treat at the optimal quality level is lower than the number of people demanding treatment. If hospitals are rewarded for the number of people they treat, they will have an incentive to treat too many. Rewarding them separately for the number of treatments and for the number of referrals for future treatment (i.e. for their demand) solves the problem. A similar argument, but for opposite reasons, applies in the case of a capacity constraint. If the hospital cannot treat more than k patients it doesn't need to increase its level of quality up to q^* in order to attract them, and will therefore choose too low a quality level[7]. Finally, a payment system like (2.7) works efficiently in all cases, i.e. whether the system is constrained or unconstrained.

In all cases the optimal setting of the price makes hospitals choose f^*. This comes from the fact that being the residual claimants of their costs gives them the incentives to cost minimisation.

2.3 The waiting lists model.

For a competitive market and with convex costs, C&M show that a linear pricing system can lead to the first best solution: hospitals have an incentive not to reduce quality levels in order to attract patients, which are their source of income. Quality does not need to be monitored or specified in the contract; all that is required is observability of demand and that this is responsive to quality[8].

[7] The generalisation to multiple quality levels is offered in C&M paper but is not dealt with here.
[8] The argument is even stronger if contracts are short term, because of a reputation effect.

The first argument is strictly linked with the discussion of the role of waiting lists, which are considered below. The second one, which is quite a strong assumption, is less critical for the case of the NHS because demand does not come directly from patients but is filtered by their GPs referrals and by the Health Authorities (as was discussed in the model above). These purchasers are in a better position to collect and process information about the quality of treatment provided by different hospitals. This is equivalent to considering health care as a search good and not an experience good (as it is for the single patient), thus avoiding much of the informational problems that are typical of the sector.

The main change brought to the C&M model is the introduction of waiting lists. These are considered as a measure of patients' demand, and thus an indirect measure of quality, taking into consideration that they negatively affect patients' utility. The model is constructed as follows.

If patients get treatment of quality q, their utility from this treatment can be defined as

$$U = q \qquad (2.8)$$

If treatment is postponed by w periods, where w is the time length of the waiting list, then (2.8) must allow for a discounting factor reflecting time preference, and for a factor reflecting the risk of a worsening of patients conditions while waiting.

Define r the time preference factor and h the conditional probability of "dying" at time t, having survived until then[9], the expected utility of a patient is

$$U = qe^{-w(r+h)} \tag{2.9}$$

Now define g as the number of people in front of a patient in the queue; its change over time will be given by

$$g = -x - hg \tag{2.10}$$

where x is the number of cases treated. Considering a steady state situation in which x (and later y, the demand) is constant, then (2.10) can be solved to obtain

$$g(t) = \left(G + \frac{x}{h}\right)e^{-ht} - \frac{x}{h} \tag{2.11}$$

Equation (2.11) means that the number of people in front of a person in the queue is a function of G (the number of people at the beginning, when $t = 0$) and it decreases with time, partly because they are treated partly because they die. When $t = w$ then $g(t) = 0$ and (2.11) can be solved with respect to w, i.e.

$$w = \left(\ln\frac{Gh + x}{x}\right)\frac{1}{h} \tag{2.12}$$

Moreover, the number of people in a queue changes over time as

$$G = y - x - hG \tag{2.13}$$

where y is the number of people joining a queue, or each hospital's demand. In steady state, when G is constant then

$$G = \frac{y - x}{h} \tag{2.14}$$

[9] In particular, F(t) is the probability distribution function of dying at any point in time;

Substituting (2.14) into (2.12) and then into (2.9) finally gives

$$U_i = q_i \left(\frac{x_i}{y_i} \right)^{\frac{r+h}{h}} \qquad (2.15)$$

Equation (2.15) is an expression of the expected utility of joining a queue at hospital i, and in equilibrium this value will be the same for every hospital.

In the social welfare function the benefit is now the expected utility of a cohort of patients asking for treatment at all hospitals at time 0; hospitals costs are discounted by a factor $e^{-rw} = (x/y)^{r/h}$ as they will be incurred only after w periods. The same discount rate r is assumed for both patients and hospitals, but these assumptions will be relaxed in the next section.

The social welfare function can be now written as

$$W = \sum_i \left\{ y_i q_i \left(\frac{x_i}{y_i} \right)^{\frac{r+h}{h}} - (1+\alpha)[C_i(x,q,f) + v_i(x,q,f)] \left(\frac{x_i}{y_i} \right)^{\frac{r}{h}} - \alpha \bar{v} \right\} \qquad (2.16)$$

In this case $\bar{v}$ is the present value of the hospitals' reservation utility.

On the hospitals' side, the objective function is like in C&M but with the cost-discounting factor. The demand function $y(x,q)$ for each hospital is derived from (2.15) as

$$y_i = \left(\frac{q_i}{U} \right)^{\frac{h}{r+h}} x_i \qquad (2.17)$$

$e^{-ht} = 1 - F(t)$ is the probability of surviving t periods, so that $h = \frac{dF(t)/dt}{1 - F(t)}$ is the conditional probability as in text.

where U is the level of utility available at other hospitals.

Substituting (2.17) into the discount factor formula finally gives

$$H = B(x, y(x,q)) - [C(x,q,f) + v(x,q,f)]\left(\frac{U}{q}\right)^{\frac{r}{r+h}} \tag{2.18}$$

which is the hospitals' objective function. Consistently with C&M results, in (2.18) the payment system $B(.)$ has been directly specified as a function of the number of cases treated and the volume of demand[10].

The simultaneous maximisation of (2.16) and (2.18) with respect to x, q and f leads to the following system of F.O.C.

$$W_x = \left[\frac{r+h}{h}q - (1+\alpha)(C_x + v_x) - \frac{r}{hx}(1+\alpha)(C+v)\right]\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \tag{2.19}$$

$$W_q = \left[x - (1+\alpha)(C_q + v_q)\right]\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \tag{2.20}$$

$$W_f = -(1+\alpha)(C_f + v_f)\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \tag{2.21}$$

$$H_x = B_x + \left(\frac{U}{q}\right)^{\frac{r}{r+h}}\left(\frac{q}{U}B_y - C_x - v_x\right) = 0 \tag{2.22}$$

$$H_q = \left[B_y \frac{hx}{(r+h)U} - (C_q + v_q) + \frac{r(C+v)}{(r+h)q}\right]\left(\frac{U}{q}\right)^{\frac{r}{r+h}} = 0 \tag{2.23}$$

$$H_f = -(C_f + v_f)\left(\frac{U}{q}\right)^{\frac{r}{r+h}} = 0 \tag{2.24}$$

[10] The consequences of other payment systems are in Appendix 2.1.

To proceed (2.19) is equated with (2.22), (2.20) with (2.23) and (2.21) with (2.24).

First of all, (2.21) and (2.24) will always be the same, that is hospitals have an incentive to maximise the effort in cost reduction, i.e. with a prospective payment system they will always choose f^*. As regards the choice of x and q, solving with respect to the price variables leads to

$$B_x = 0$$
$$B_y = \frac{U}{1+\alpha}\frac{r+h}{h} - \frac{r(C+v)U}{hqx} \qquad (2.25)$$

Rearranging the equations[11], another way to express the result for B_y is

$$B_y = \frac{U_q y}{y_q} - \frac{e^{-rw}\left[\alpha(C_q + v_q) + \frac{C+v}{q}\frac{r}{r+h}\right]}{y_q} \qquad (2.26)$$

Equation (2.26) is the same price per patient demanding treatment as in C&M (see equation (2.7)) except for the presence of the discount factor and its sensitivity to quality changes. As could be expected, the optimal price, and hence the marginal benefit for the hospital, equals the marginal social benefit, net of a portion α of the marginal cost. This result indicates that it is optimal to pay the hospital a fixed sum T and then a fixed price for every person asking for treatment by entering its waiting list, and a zero price for the number of people actually treated.

This apparently counterintuitive result stems from the way demand was modelled. On the one hand, hospitals have an incentive to increase the quality

level of their service: they are paid on the basis of their demand and, given the assumptions, a higher quality attracts more people. On the other hand, however, more people demanding treatment translate into a longer waiting time, and this in turn acts negatively on demand. In order to counteract this effect hospitals will decide the number of people they want to treat, as the more patients are treated the shorter is the waiting time. In other words the adjustment towards equilibrium can be described as an adjustment towards an optimal waiting time; rewarding hospitals on the basis of their demand gives them immediate and direct incentives over q and also indirect incentives over x.

If only a price per case were paid, and not a price per person asking for treatment, then hospitals would treat the right number of people but at a too low quality level. If in order to give them an incentive to increase quality they were partly reimbursed for their costs, then they would not make the right effort in cost reduction. The results for these two last cases are derived in Appendix 2.1.

The payment of both a positive price per case and per demand would lead to the treatment of too many patients. This result will become clear in the next section, when the social cost of having people waiting and eventually dying is considered.

[11] The result is obtained if in the F.O.C of the welfare function the derivatives of U are expressed as U_x and U_y, and in the hospital's F.O.C the same is done for the derivatives of y (y_x and y_q).

2.4 Extensions to the model.

The model developed in Section 2.3 is based on several assumptions, some of which have already been discussed. Three of these assumptions will now be relaxed, to see if and how they affect the results obtained. In particular, the following changes will be introduced:

1) Instead of considering that hospitals and patients have the same discount factor r for time preference, a different factor is introduced to discount patients' utility when treatment is postponed.
2) The assumption of perfect competition is relaxed to see how the model behaves in the case of an oligopoly of hospitals.
3) Finally, the implicit assumption that social and private costs are the same is relaxed and the "cost of the loss" of patients getting worse or dying while waiting is introduced in the social costs.

2.4.1 Different discount factor.

It might be the case that people's discount factor for time preference is different from that of hospitals. Calling the former δ and the latter r, the objective functions can be re-written as

$$W = \sum_i \left[y_i q_i \left(\frac{x_i}{y_i} \right)^{\frac{\delta+h}{h}} - (C+v) \left(\frac{x_i}{y_i} \right)^{\frac{r}{h}} (1+\alpha) \right] \tag{2.27}$$

$$H = B(x.y) - (C+v) \left(\frac{U}{q} \right)^{\frac{r}{\delta+h}} \tag{2.28}$$

Proceeding as before, the following results are obtained

$$B_x = 0 \tag{2.29}$$

$$B_y = \frac{U}{1+\alpha}\frac{\delta+h}{h} - \frac{r(C+v)}{hx}\left(\frac{U}{q}\right)^{\frac{r+h}{\delta+h}} \tag{2.30}$$

which is the same as

$$B_y = \frac{U_q y}{y_q} - \frac{e^{-rw}[\alpha(C_q+v_q)+\frac{r}{\delta+h}\frac{C+v}{q}}{y_q} \tag{2.31}$$

The result and its rationale are therefore the same as before. It is still optimal to reward hospitals on the basis of their demand, paying them a fixed price per person entering the queue; this is calculated in the same way as before. In this case, as can be seen from (2.30) or (2.31), an increase in δ will have a positive effect on the price B_y, meaning that a stronger dislike for waiting leads to a higher payment per patient to the hospital. Unfortunately, the calculation of the effect of δ on the equilibrium is inconclusive. Intuitively one might expect the higher marginal revenue to translate into an incentive to increase x and q, in order to increase y, possibly leading to higher equilibrium values of both quality and quantity. However, the outcome where either only x or q is increased cannot be ruled out.

2.4.2 Oligopoly of supply.

The case is now considered of an oligopolistic market with a finite number m of homogeneous hospitals. Whereas the homogeneity assumption might still be simplifying, the hypothesis of a finite number of hospitals is quite realistic,

especially for local market conditions. The case is modelled as a Cournot-Nash game, in which hospitals decide their quality-quantity mix at the same time, taking the others' decisions as given.

This interaction between hospitals can be introduced in the model via a respecification of their demand function. In a symmetric case with m hospitals the demand of each hospital in equation (2.17) becomes

$$y_i = \left(\frac{q_i}{U}\right)^{\frac{h}{r+h}} x_i \quad \forall i \qquad (2.32)$$

$$i = 1,\ldots,m$$

and the total market demand is

$$Y = \sum_i y_i$$

As a symmetric equilibrium is assumed, (2.32) can be equivalently rewritten as

$$y_i = \frac{\left(\frac{q_i}{U}\right)^{\frac{h}{r+h}} x_i}{\sum_j \left(\frac{q_j}{U}\right)^{\frac{h}{r+h}} x_j} Y \quad \forall i \qquad (2.33)$$

Equation (2.33) shows that every decision made by hospital i about its quality or quantity level will affect it both directly, through the effect on its demand, and indirectly, through the effect on the others' market shares. Differentiating (2.33) with respect to x and q gives

$$y_x = \frac{y}{x}\left(\frac{m-1}{m}\right)$$

$$y_q = \left(\frac{h}{r+h}\right)\left(\frac{y}{q}\right)\left(\frac{m-1}{m}\right)$$

The same procedure as before is then used (maximise the social welfare function and the hospitals' utility function and equate their F.O.C.) but using the above expressions for the marginal variation of demand with respect to quantity and quality, because this specification explicitly takes into consideration the number of firms and their market shares. This gives the following results:

$$B_x = 0$$

$$B_y = \frac{m}{m-1}\left(\frac{U}{q}\right)^{\frac{r}{r+h}}\left[\frac{x}{(1+\alpha)}\frac{(r+h)}{xy} - \frac{r(C+v)}{hy}\right] \qquad (2.34)$$

This is exactly the same result of (2.25) multiplied by *m/(m-1)*.

Thus again, in the case of an oligopoly it is also optimal to pay hospitals a zero price for the number of cases treated and to reward them instead on the basis of their demand. It is therefore only the price per entrant to be influenced by the number of hospitals on the market. As can be seen by differentiating (2.34) with respect to m, the higher is this number the lower is the price and viceversa, provided that $m \geq 2$. That is, the result does not hold in the case of a monopolistic provider. This is quite intuitive, as the whole model works on the adjustments of demand. In the case of a monopoly demand is fixed, as patients do not have any choice as of where to go to receive treatment. Formally this means that both y_x and y_q are zero. From the objective function of the hospital it can be easily seen that it would provide too low a quality level, even though a positive price per case could induce it to treat the right number of patients.

2.4.3 The social "cost of the loss".

So far, the only costs that have been considered are the production cost of the hospital and its disutility. It is realistic however to consider that, from a social welfare point of view, the fact that people might get worse while they wait and eventually die has a cost, other than their own cost from waiting. In other words, not only people, but also society, are worse off the longer they have to wait; the very events of getting worse and eventually dying carry a cost, that includes rather unmeasurable as well as more measurable elements (for example the payment of benefits to people when they cannot work). Call this "cost of the loss" γ. This has to be introduced in the social welfare function that the purchaser is assumed to be maximising. This means that social and private costs are now different.

Recall from footnote (9) that

$F(t) = 1 - e^{-ht}$

is the cumulative distribution of the probability of dying at time t. If γ is the social cost of the loss, the expected social cost today of having y persons getting worse between now and w is given by

$$yh\gamma\int_0^w e^{-(r+h)t}dt \tag{2.35}$$

From (2.35) the value of the social cost of the loss is

$$y\gamma\frac{h}{r+h}\left[1-\left(\frac{x}{y}\right)^{\frac{r+h}{h}}\right] \tag{2.36}$$

A new expression for the welfare function can now be rewritten as

$$W = \sum_i \left\{ y_i q_i \left(\frac{x_i}{y_i} \right)^{\frac{r+h}{h}} - (1+\alpha) \left[(C_i + v_i) \left(\frac{x_i}{y_i} \right)^{\frac{r}{h}} - y\gamma \frac{h}{r+h} \left(1 - \left(\frac{x}{y} \right)^{\frac{r+h}{h}} \right) \right] \right\} \quad (2.37)$$

Applying the same procedure as in all previous cases, the result is now different, i.e.

$$B_x = \gamma$$
$$B_y = \frac{U}{1+\alpha} \frac{r+h}{h} - \frac{r(C+v)U}{hqx} \quad (2.38)$$

The value of the price per person asking for treatment (B_y) is the same as in (2.25), but in this case a positive price per case (B_x) is required, which is equal to the social cost of the loss itself.

To interpret this result it is necessary to analyse what effect the introduction of γ has had on the equilibrium. To do so, let's consider the following equations from the F.O.C. for welfare maximisation:

$$W_x = \frac{r+h}{h} q - (1+\alpha)(C_x + v_x) - (1+\alpha) \left(\frac{r}{h} \frac{C+v}{x} - \gamma \right) = 0 \quad (2.39)$$

$$W_q = x - (1+\alpha)(C_q + v_q) = 0 \quad (2.40)$$

The effect that the introduction of γ has had on the optimal values of x and q can be calculated by totally differentiating (2.39) and (2.40), which results in

$$\begin{aligned} &\left[-(1+\alpha)(C_{xx} + v_{xx}) - (1+\alpha) \frac{r}{h} \left(\frac{(C_x + v_x)x - (C+v)}{x^2} \right) \right] dx + \\ &+ \left[\frac{r+h}{h} - (1+\alpha)(C_{xq} + v_{xq}) - (1+\alpha) \frac{r}{h} \frac{(C_q + v_q)}{x} \right] dq = -(1+\alpha) d\gamma \end{aligned} \quad (2.41)$$

$$[1 - (1+\alpha)(C_{xq} + v_{xq})] dx + [-(1+\alpha)(C_{qq} + v_{qq})] dq = 0 \quad (2.42)$$

In matrix notation this can be rewritten as

$$H\begin{pmatrix} dx \\ dq \end{pmatrix} = \begin{bmatrix} -(1+\alpha)d\gamma \\ 0 \end{bmatrix} \tag{2.43}$$

where H is the Hessian matrix of the second order derivatives of the welfare function. From (2.43) finally get

$$\frac{dx}{d\gamma} = \frac{(1+\alpha)^2(C_{qq}+v_{qq})}{|H|} \tag{2.44}$$

$$\frac{dq}{d\gamma} = \frac{(1+\alpha)[1-(1+\alpha)(C_{xq}+v_{xq})}{|H|} \tag{2.45}$$

From the second order conditions the determinant of the Hessian matrix is known to be positive, so that

$$\frac{dx}{d\gamma} > 0 \quad \text{always}$$

$$\frac{dq}{d\gamma} > 0 \quad if \quad (C_{xq}+v_{xq}) < \frac{1}{1+\alpha}$$

$$\frac{dq}{d\gamma} < 0 \quad if \quad (C_{xq}+v_{xq}) > \frac{1}{1+\alpha}$$

This means that the introduction of γ always produces an increase in the number of patients to treat, whereas the effect on the quality level depends on the size of the cross derivative of the hospital cost and disutility function. To shed more light on the above, consider the cross derivative W_{qx} as calculated from (2.40); it can be seen that when

$$(C_{xq}+v_{xq}) < \frac{1}{1+\alpha} \tag{2.46}$$

then $W_{xq}>0$, that is the marginal benefit of quality increases with the number of cases and the sign of $\frac{dq}{dx}$ (obtained via the total differentiation of (2.40)) is positive. The opposite holds in the case

$$(C_{xq} + v_{xq}) > \frac{1}{1+\alpha} \qquad (2.47)$$

This means that if (2.46) is true then an increase in the number of cases makes an improvement in quality more valuable to society. The increase in the optimal number of cases brought about by γ will therefore lead to a higher level of quality. The opposite holds in the case when (2.47) is true. Another way of looking at this is via the behaviour of the average cost function. From the convexity assumptions of the total cost function made at the beginning, the average cost per case $(C+v)/x$ is an increasing function of both x and q, that is the average cost of treating a patient gets higher the more patients are treated and/or the higher is the quality level.

The implication of (2.46) is that the average cost will tend to increase with quality less when the hospital treats more people. If instead (2.47) is true then the increase in the average cost per case will be higher the higher is the number of patients treated. In the first case, a social welfare function that includes the cost of the loss will be maximised by increasing both quantity and quality; in the second case, the increase in the optimal number of cases will lower the socially optimal quality level.

Even if this last hypothesis might seem more reasonable to expect, the opposite cannot be ruled out a priori. This could be the case for example if in order to increase the number of patients to treat the hospital has to buy and use some particular equipment, or hire some specialised staff, and this has beneficial effects on the quality level of treatment that can be increased at a lower extra cost.

In conclusion, the social cost of the loss introduces a difference between the social and the private cost functions, and the optimal number of cases is increased. As a consequence, a positive price per case is now required to give hospitals correct incentives. The optimal price per entry in the waiting list is determined in the same way as before, and its value will depend on whether the optimal quality level has increased or decreased.

2.5 Conclusions.

The aim of this chapter was to develop a model for the identification of an efficient contract for hospital services, that is a contract that gives hospitals the correct incentives to minimise their costs and to choose the right quality and quantity of treatment. In particular, the work by C&M has been used as a basis to develop a model that explicitly takes into consideration the existence of waiting time, and the fact that this affects patients utility. The model assumes that quality is not enforceable by contract, nor can be easily or costlessly monitored. However, it is assumed that purchasers, the agents of the demand, correctly perceive it and thus respond to it. Costs are convex and known to the purchaser.

In this framework the optimal solution is a prospective payment system that rewards hospitals on the basis of their demand rather than their output. A positive price per case would result in either the treatment of too many people (if associated to a price per demand) or in the choice of too low a quality level (if no price per demand were paid). A price per person demanding treatment gives direct incentives to quality and indirect incentives to the number of patients to treat. The equilibrium can be described therefore as an adjustment towards an optimal waiting time. The same result holds true when a different discount factor for time preference is considered, and in the case of an oligopolistic market. In the first case, the price is shown to be an increasing function of how much people dislike waiting, in the second case a decreasing function of the number of providers.

The additional payment of a price per case has been shown to be necessary when a difference between the social and the private cost functions is modelled. The social cost of having people waiting and eventually getting worse and dying is considered. This translates into a higher optimal number of cases treated, which in turn requires a positive price per case to work as an incentive.

The results of the model rely on the assumptions made about the observability of costs and the responsiveness of demand to quality. As seen in Section 2.1, the relaxation of either of these assumptions usually changes the results pointing towards more complex payment systems, some of which might include forms of

cost reimbursement and not necessarily lead to the first best solution. Such extensions to the present model are left for future research.

A general conclusion that can be drawn from the theoretical analysis is that if an optimal contract can be devised, this is far from being a simple task. The reality of the actual contractual agreements confirms this complexity. Vast amount of resources, including the recruitment of specialised staff, were used by both purchasers and providers (Robinson and Le Grand, 1994; Fattore, 1999). Especially at the beginning, extensive use was made of simple block contracts, then substituted by "sophisticated" block contracts, which would define upper and lower thresholds of activity, agreements on the monitoring of performance and the possibility of renegotiations. Progressively the system moved towards forms of prospective payment, i.e. cost-per-case and cost-and-volume contracts. Broadly speaking these would reward the hospitals on the basis on the number of cases treated, although their level of detail would be extremely high. The properties of these contracts have been analysed by the theory, which pointed to some possible drawbacks especially (but not exclusively) as regards the quality of the service. Their major advantage would lie in their cost saving, or waste reducing, incentives, which were also one of the main principles guiding the reform. This particular efficiency issue can be empirically analysed, and this is the aim of the second part of the thesis.

APPENDIX 2.1

The results of the model developed in Section 2.3 showed as an optimal pricing rule to reward the hospital for every person demanding treatment. In this appendix, the effects of choosing a different pricing rule are shown. The case of a pure cost reimbursement system is quite straightforward and is therefore omitted. Two other possible cases are instead presented: the payment of a price per case only, and the case of a price per case with partial cost reimbursement. As will be shown, both lead to sub-optimal results.

CASE 1: PAYMENT OF A POSITIVE PRICE PER CASE AND A ZERO PRICE PER DEMAND.

If the payment system to the hospital was set as a function of the number of cases treated only, that is B_y=0, then the hospital would treat the right number of cases but choose too low a quality level.

The F.O.C from the maximisation of (2.16) and (2.18) become

$$W_x = \left[\frac{r+h}{h}q - (1+\alpha)(C_x + v_x) - \frac{r}{hx}(1+\alpha)(C+v)\right]\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \qquad \text{(A1.1)}$$

$$W_q = \left[x - (1+\alpha)(C_q + v_q\right]\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \qquad \text{(A1.2)}$$

$$W_f = -(1+\alpha)(C_f + v_f)\left(\frac{x}{y}\right)^{\frac{r}{h}} = 0 \qquad \text{(A1.3)}$$

$$H_x = B_x - \left(\frac{U}{q}\right)^{\frac{r}{r+h}}(C_x + v_x) = 0 \qquad \text{(A1.4)}$$

$$H_q = -\left(\frac{U}{q}\right)^{\frac{r}{r+h}}\left[(C_q + v_q) - \frac{r(C+v)}{(r+h)q}\right] = 0 \qquad \text{(A1.5)}$$

$$H_f = -(C_f + v_f)\left(\frac{U}{q}\right)^{\frac{r}{r+h}} = 0 \qquad \text{(A1.6)}$$

Equations (A1.3) and (A1.6) are always the same, so the hospital is putting the required level of effort in cost reduction.

Equation (A1.1) implies that

$$(C_x + v_x) = \frac{r+h}{h}\frac{q}{1+\alpha} - \frac{r(C+v)}{hx}$$

and equation (A1.4) implies that

$$B_x = \left(\frac{U}{q}\right)^{\frac{r}{r+h}}(C_x + v_x)$$

One can therefore set

$$B_x = \left(\frac{U}{q}\right)^{\frac{r}{r+h}}\left(\frac{r+h}{h}\frac{q}{1+\alpha} - \frac{r(C+v)}{hx}\right)$$

and this guarantees that the hospital(s) will choose the optimal value of (C_x+v_x) and then choose x^*.

However, from equation (A1.2)

$$(C_q + v_q) = \frac{x}{1+\alpha}$$

and from equation (A1.5)

$$(C_q + v_q) = \frac{r}{r+h}\frac{C+v}{q}$$

For the solution to be optimal it must therefore be true that

$$\left(\frac{x}{1+\alpha}\right)^* = \frac{r}{r+h}\frac{C+v}{q}$$

From equation (A1.1)

$$\left(\frac{1+\alpha}{x}\right) = \frac{(r+h)q}{r(C+v)} - \frac{(1+\alpha)(C_x + v_x)}{r(C+v)}$$

and therefore

$$\frac{x}{1+\alpha} = \frac{r(C+v)}{(r+h)q - (1+\alpha)(C_x + v_x)} > \frac{r(C+v)}{(R+h)q}$$

i.e. the socially optimal quality level is lower than that chosen by the hospital.

CASE 2: PAYMENT OF A PARTIAL COST REIMBURSEMENT OF THE TYPE $B(C) = B_x + \varphi C$.

As the payment of a price per case as in Case 1 leads to too low a quality level, an alternative could be to partially reimburse the hospitals for their total costs in order to give them incentives to increase the quality level of the service. However, this solution does not lead to the first best either.

If the hospital were paid a fixed price per case and were reimbursed of a portion φ of its total costs it would not choose the right level of effort in cost reduction.

Let's call for simplicity the discount factor z, so that

$$z = \left(\frac{U}{q}\right)^{\frac{r}{r+h}} = \left(\frac{x}{y}\right)^{\frac{r}{h}}$$

The F.O.C can be written now as:

$$W_z = \left[\frac{r+h}{h}q - (1+\alpha)(C_x + v_x) - \frac{r}{hx}(1+\alpha)(C+v)\right]z = 0 \qquad \text{(A2.1)}$$

$$W_q[x - (1+\alpha)(C_q + v_q)]z = 0 \qquad \text{(A2.2)}$$

$$W_f = -(1+\alpha)(C_f + v_f)z = 0 \qquad \text{(A2.3)}$$

$$H_x = B_x + \varphi C_x - z(C_x + v_x) = 0 \qquad \text{(A2.4)}$$

$$H_q = \varphi C_q - z\left[(C_q + v_q) - \frac{r(C+v)}{q(r+h)}\right] = 0 \qquad \text{(A2.5)}$$

$$H_f = \varphi C_f - (C_f + v_f)z = 0 \qquad \text{(A2.6)}$$

Whatever value is calculated for B_x it is obvious from the comparison of (A2.3) and (A2.6) that the hospital will not choose the optimal effort in cost reduction unless $\varphi=0$.

In fact, from (A2.3) the optimal solution implies that

$v_f = - C_f$

which is >0 as v is increasing in f and C is always decreasing in f.

From (A2.6) instead

$$v_f = \frac{\varphi - z}{z} C_f \qquad \text{(A2.7)}$$

If $\varphi>z$ then (A2.7)is <0, implying a negative effort in cost reduction. If $\varphi<z$ then (A2.7) is positive, but as $(\varphi-z)/z>-1$ not enough effort is put anyway.

CHAPTER 3

THE MEASUREMENT OF PRODUCTIVE EFFICIENCY

The general aim of this chapter is to provide an explanation of the concepts of efficiency and productivity, and to review the main literature contributions to their measurement.

Productive, or technical, efficiency is defined with respect to inputs (or outputs) *levels*, whereas cost efficiency is the result of a process of cost minimisation and it embodies also the concept of allocative efficiency which is related to inputs *proportions*, given the inputs prices. Because this thesis performs an analysis of productive efficiency, for reasons of space the focus of this chapter will be on technical efficiency only.

The chapter is structured as follows. The economic concepts of efficiency and productivity are discussed in Section 3.1. Sections 3.2 to 3.4 explain the different measurement techniques: data envelopment analysis is the object of Section 3.2, the econometric estimation of frontiers is in Section 3.3, with the particular case of panel data in Section 3.3.1, and Section 3.4 describes the issue of measuring technical progress and shifts of the frontier. Finally, the applications to the case of hospital services are reviewed in Section 3.5.

A general comparison between the methodologies is done in Chapter 6 with the conclusions, after the analysis of Chapters 4 and 5.

3.1 The microeconomic concepts of technical efficiency and productivity.

Productivity and efficiency are two entwined concepts; the former is a general measure of the ratio of output(s) to input(s), the latter entails a comparison of the actual (observed) ratio to an optimal one which is usually referred to as the "frontier".

Following Koopmans (1951), in a multiple outputs - multiple inputs case technical efficiency is a situation such that it is impossible to increase even just one output without either decreasing at least another output or increasing at least one input; or, viceversa, it is impossible to decrease even just one input without either increasing another input or decreasing at least one output. In other words, it is the maximum attainable output given a set of inputs, or the minimum level of inputs required to produce a given level of output.

More formally let's define the following[1].

If a vector $x \in R_+^K$ of inputs is used to produce a vector $y \in R_+^M$ of outputs, then

$$L(y) = \{x : (y, x) \; is \; feasible\} \tag{3.1}$$

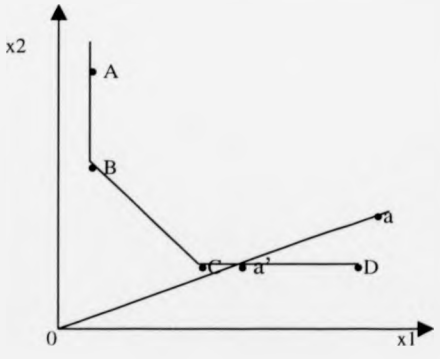
is the *input requirement set,* and

$$\begin{aligned} IsoqL(y) &= \{x : x \in L(y), \lambda x \notin L(y), \lambda \in [0,1)\} \\ Eff\, L(y) &= \{x : x \in L(y), x' \notin L(y), x' \leq x\} \end{aligned} \tag{3.2}$$

[1] See for example Lovell C. A. K. (1993).

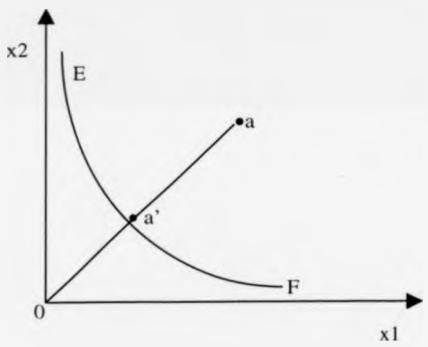
are respectively the *isoquant* and the *efficient subset*. The input requirement set identifies a feasibility set for the inputs, that is all the input levels that are sufficient, though not necessarily efficient, to produce the output vector *y*. In other words, it represents the production technology. The isoquant is the boundary of the input requirement set, and is defined in terms of radial contraction of the input points within it. When the production technology is represented by a well behaved, continuously differentiable function the isoquant is the same as the efficient subset (the frontier), and it represents the minimum input level necessary to produce a given output level. If that is not the case, like for example if the frontier is piece-wise linear, the isoquant and the efficient subset do not coincide. This is shown graphically in Fig. 3.1 and 3.2, for the two inputs- one output case.

Fig.3.1: Piece-wise linear frontier, input minimisation case with one output and two inputs (x1 and x2).



On the piece-wise linear frontier of Fig.3.1 the isoquant ABCD[2] does not coincide with the efficient subset BC: a point like *a'* in fact cannot be deemed efficient as the same level of output could be produced by reducing the level of input *x1*, as for point *C*. The difference in the level of *x1* between *a'* and *C* is called an input slack.

Fig.3.2: Continuously differentiable frontier, input minimisation case with one output and two inputs (x1 and x2).



In the case of Fig.3.2, instead, the frontier is continuously differentiable and the input set is strictly convex so that the isoquant and the efficient subset are the same line EF. A case like that of Fig. 3.2 represents a typical well-behaved production function, like most theoretical neoclassical production functions, with positive marginal productivities of the inputs. The horizontal facets of the isoquant in Fig. 3.1 instead correspond to marginal productivities equal to zero (inputs strong disposability). The case of negative marginal productivities (or

[2] For clarity of explanation: points A and D are not assumed to be actual observations, but are mentioned in the figure to identify the isoquant.

weak disposability, an assumption more rarely allowed for, also known as inputs congestion) would translate in positively sloped facets[3].

In both figures the input requirement set $L(y)$ is the area on and above the isoquant.

From an output perspective, similarly

$$P(x) = \{y : (x, y) is\ feasible\} \tag{3.3}$$

is the *output set*, i.e. all the levels of output that can be produced using a given level of inputs, whether efficient or not.

$$\begin{aligned} IsoqP(x) &= \{y \in P(x), \vartheta y \notin P(x), \vartheta(1, \infty)\} \\ EffP(x) &= \{y \in P(x), y' \notin P(x), y' \geq y\} \end{aligned} \tag{3.4}$$

are the isoquant and the efficient subset. The isoquant is the boundary of the output set, and it is defined in terms of radial expansions of the output points within it.

Again, in the case of a well behaved, continuously differentiable production technology the isoquant and the efficient subset coincide, and in this case they represent the maximum level of output which can be produced from a given level of inputs.

A graphical representation is offered in Fig.3.3 and 3.4 for the two outputs-one input case.

[3] A more detailed discussion of the disposability assumptions is in Cubbin and Ganley, 1992.

Fig.3.3: Piece-wise linear frontier, output maximisation case with one input and two outputs (y1 and y2).

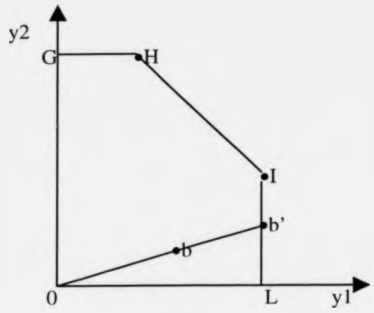
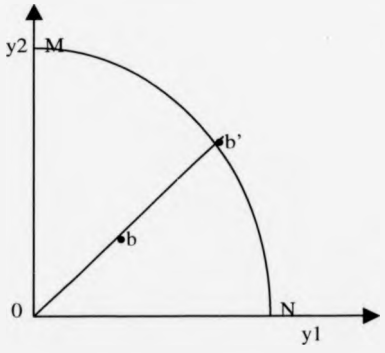


Fig.3.4: Continuously differentiable frontier, output maximisation case with one input and two outputs (y1 and y2).



Similarly to Fig.3.1 and 3.2, on the piece-wise linear frontier the efficient subset HI is not the same as the whole isoquant GHIL, and a point as *b'* is not efficient because more of output *y2* could be produced with the given input level, as for point *I*. The difference in the level of *y2* between *b'* and *I* is called an output slack.

In all cases the output set *P(x)* is the area on and below the isoquant.

The above definitions thus identify the efficient frontier, whether that is expressed in terms of output maximisation or of inputs minimisation. Efficiency of a particular firm can be defined in terms of its distance from that frontier, i.e. by means of a distance function. The distance function was defined by Shephard (1953, 1970), and it is equivalent to the definition of efficiency of Debreu (1951) and Farrell (1957)[4]; it is defined as the equiproportionate increase (in outputs) or decrease (in inputs) necessary to reach the frontier, i.e. it is a *radial* measure.

Using Shephard's notation, in the input minimisation perspective the input distance function is

$$D_I = \max\left\{\lambda : \frac{x}{\lambda} \in L(y)\right\} \qquad (3.5)$$

where $D_I \geq 1$. If $D_I = 1$ the observed firm is efficient, as it lies on the frontier. A value of $D_I > 1$ indicates inefficiency, measured by the radial contraction $1/\lambda$ necessary to reach the frontier, which is equivalent to a $[1-(1/\lambda)]*100$ percentage change. In terms of Fig.3.1 and 3.2, it corresponds to the ratio $0a/0a'$. The input distance function is linearly homogeneous of degree +1 and weakly monotonically increasing in inputs, and is invariant to changes in the units of measurement[5].

[4] In particular, the Debreu-Farrel efficiency measure is the reciprocal of Shephard's distance function.

[5] These properties of the distance function are discussed, among others, by Shephard (1970) and Fare and Lovell (1978). For general reference see Fried, Lovell and Schmidt, 1993.

Similarly, in the case of output maximisation the output distance function is given by

$$D_o = \min\left\{\vartheta : \frac{y}{\vartheta} \in P(x)\right\} \tag{3.6}$$

where $0<D_o\leq 1$. Again, if $D_o = 1$ the observation lies on the frontier, if $D_o < 1$ it lies below it and a radial expansion of $1/\vartheta$ of the outputs is necessary to reach it, equivalent to a $[(1/\vartheta)-1]*100$ percentage change. Looking at Fig.3.3 and 3.4 this corresponds to the ratio $0b/0b'$.The output distance function is homogeneous of degree +1 and weakly monotonically increasing in outputs, and is invariant to changes in the units of measurement.[6]

One thing has to be noticed before concluding. As already noticed, if the frontier is a continuously differentiable function then the isoquant and the efficient subset coincide, and in that case the radial efficiency measure coincides with Koopmans' definition of technical efficiency. This is not true when the frontier is not continuously differentiable. In a case like that of point *a* of Fig. 3.1 (point *b* in Fig. 3.3), the radial measure of efficiency identifies point *a'* (*b'*) which is inefficient because of the input (output) slack. A proper measure of (in)efficiency therefore requires some kind of adjustment. This problem and its solutions will

[6] In the case of cost efficiency, the frontier is the minimum cost for producing a given level of output, and the distance from that frontier measures the excess cost of the firm. This includes both technical inefficiency (waste in inputs) and allocative inefficiency (wrong inputs proportions), which can then be disentangled. However, as said at the beginning, no more detail is given to the characteristics of cost frontiers for reasons of space, given that the thesis estimates production frontiers anyway.

be discussed in the section on Data Envelopment Analysis (DEA), because the frontier that DEA estimates is piece-wise linear.

3.2 Data Envelopment Analysis.

There are two main approaches to the estimation of frontiers and the measurement of efficiency: the linear programming, non-parametric techniques of DEA and the econometric, parametric techniques. The former are usually deterministic, though some contributions to a stochastic version have been recently developed by the literature[7]. The latter can be both deterministic and stochastic, but as the lack of statistical noise is a strong limitation they are more commonly used in their stochastic version. As a consequence of these characteristics, they have opposite advantages and disadvantages, as will be discussed in the conclusions.

DEA[8] is a mathematical programming technique that was developed by Charnes, Cooper and Rhodes in 1978 (the CCR model) to measure efficiency in the non-profit sector. Other models have followed the original CCR paper, and they differ with respect to the envelopment surfaces used (the way in which the frontier is identified), the orientation or focus and so on. The characteristic of DEA models is that the frontier is calculated using linear programming, i.e. it is the result of the linear combinations (envelopment surfaces) of those observations (DMUs, Decision Making Units) that use comparatively less inputs to produce

[7] See for example Olesen and Petersen, 1995.
[8] Comprehensive reviews on DEA can be found for example in Fried, Lovell and Schmidt (1993); Charnes, Cooper, Lewin and Seiford (1994); Coelli, Rao and Battese (1998); Cooper, Seiford and Tone (2000). A discussion of its application in the public sector in Ganley and Cubbin (1992).

comparatively more outputs. This piece-wise linear frontier reflects the best observed practice in the sample because the models are (mainly) deterministic.

A first general distinction to be drawn is between "orientated" DEA models, and "additive" models. The former provide a radial measure of inefficiency and require one to choose between an output and an input orientation; the latter calculate a summary additive measure of inefficiency and do not require any choice in orientation.

A simple orientated DEA model is the one first presented in CCR, and it can be depicted as follows. Assume there are $i=1,....,N$ DMUs which use the inputs $k=1,.....,K$ to produce $m=1,.....,M$ outputs. Under the assumptions of constant returns to scale, convexity of the feasible set and strong disposability of inputs and outputs[9], in an output maximisation perspective the efficiency of each of the N DMUs (in turn denominated as DMU_0) is calculated as:

$$\begin{aligned} &\max\left[\theta_o + \varepsilon\left(\sum_{k=1}^{K} s_k^- + \sum_{m=1}^{M} s_m^+\right)\right] \\ &s.t. \quad \sum_{i=1}^{N} x_{ki}\lambda_i = x_{ko} - s_k^- \\ &\sum_{i=1}^{N} y_{mi}\lambda_i = \theta_o y_{mo} + s_m^+ \\ &\lambda, s_k^-, s_m^+ \geq 0 \quad \forall i,k,m \end{aligned} \tag{3.7}$$

where θ_0 is the radial contraction for DMU_0, λ is a $1xN$ vector of weights, s_k^- and s_m^+ are respectively the inputs and outputs slacks and ε is a positive, infinitely small number.

The calculation is performed N times, once for each DMU. The efficiency of each unit is calculated in two stages as a radial measure θ plus the necessary adjustments in inputs and/or outputs s_k^- and s_m^+. In the first stage, the optimal value of θ is calculated, and in the second stage the sum of any remaining slacks is maximised to properly identify the "efficient comparator" of the unit under observation.

Using the notation in (3.6),

$\theta = 1/D_o \geq 1$

i.e. for an efficient unit $\theta = 1$ and for an inefficient one $\theta > 1$. The presence of the slack variables is a consequence of the fact that on the DEA piece-wise linear frontier, the isoquant and the efficient subset do not coincide. Looking at Fig.3.3 the radial expansion θ to point b identifies the projected point b', but a further increase in output y_2 is necessary to reach the Koopmans' efficient point 1.

In an input- minimisation perspective the DEA envelopment problem is:

$$\begin{aligned} &\min\left[z_o - \varepsilon\left(\sum_{k=1}^{K} s_k^- + \sum_{r=1}^{m} s_m^+ \right) \right] \\ &s.t. \quad \sum_{i=1}^{N} y_{mi}\lambda_i = y_{mo} + s_m^+ \\ &\qquad \sum_{i=1}^{N} x_{ki}\lambda_i = z_o x_{ko} - s_k^- \\ &\qquad \lambda, s_k^-, s_m^+ \geq 0 \quad \forall i,k,m \end{aligned} \tag{3.8}$$

In this case, the radial measure is given by z, and using the notation in (3.5)

$z = 1/D_I \leq 1$

[9] These assumptions can be relaxed (references in footnote 7).

i.e. DEA calculates the Debreu-Farrel measure of inefficiency, which is the inverse of Shephard's distance function. In terms of Fig.3.1, the DEA radial contraction corresponds to the ratio *0a'/0a*. Again, a total measure of inefficiency for each unit is calculated in two stages.

Other models have followed the original CCR paper. Banker, Charnes and Cooper (1984; BCC hereinafter) introduced the possibility of variable returns to scale (VRS). This is carried out by adding to (3.8) the additional constraint that

$$\sum \lambda_i = 1$$

which defines a convex hull, as opposed to the conical hull of the constant returns to scale hypothesis[10] (CRS).

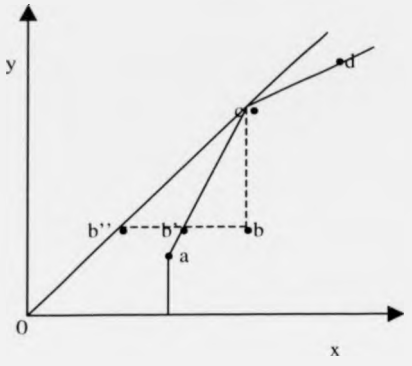
The hypothesis of non-increasing returns to scale (i.e. only constant or decreasing) can be modelled by imposing the restriction

$$\sum \lambda_i \leq 1$$

Fig.3.5 shows the difference between a VRS and a CRS frontier for the one output-one input case. The line *0c* represents the CRS frontier, whereas *acd* is the VRS frontier.

[10] This technically means that if no restriction is imposed to the λ then all supporting hyperplanes can pass through the origin.

Fig. 3.5: DEA frontier with constant and variable returns to scale, in the case of one input x and one output y.



By looking at the figure the following characteristics can be noticed:

1. Fewer units are efficient on the CRS frontier: in the figure, point *c* is sufficient to identify it, whereas on the VRS frontier more units are efficient (points *a*, *c*, and *d*). The VRS frontier envelops the data more tightly, which implies more units will be qualified as efficient.
2. As a consequence of the above, an inefficient unit under VRS has a lower inefficiency score than under CRS. In the figure, in an input perspective the distance of a point like *b* from the CRS frontier is the segment *bb''*, which is bigger than the distance from the VRS frontier *bb'*. The segment *b'b''* measures what is defined as scale inefficiency: at the optimal scale of operation returns to scale cannot be increasing or decreasing, so the difference (if any) between the CRS frontier and the VRS frontier is inefficiency of scale[11].
3. Under VRS the measure of inefficiency (but not the ordering) varies with the orientation of the model. Taking again a point like *b*, the distance from the

VRS frontier in an output perspective is the segment *bc*, which is bigger than the input measure *bb'*; in the case of CRS by definition the segments *bc* and *bb"* are the same length. The orientation of the model therefore matters when using a VRS perspective, and it raises questions very similar to those of endogeneity and exogeneity of variables in a parametric context (which variables the firm actually controls).

The difference between the VRS and the CRS efficiency scores shows the existence but not the nature of scale inefficiency. This information can be obtained by comparing the efficiency scores under VRS and NIRS (not in the figure)[12].

Both the CCR and BCC papers, as well as other "orientated" models, calculate the slack variables after the calculation of the radial measure by the two-stage process outlined before.[11] This method however has its own limitations[14], among which is the fact that it is not invariant to the units of measurement. An alternative to the radial measure of inefficiency is provided by the "additive" model[15], which replaces the radial measure with an additive measure that sums only the slacks.

---

[11] The optimal scale of production should be the one that minimises the long run average cost curve, and at that level returns to scale are constant.
[12] As the estimations of Chapter 4 will be done under the CRS assumption (for reasons of degrees of freedom) no further detail is given here on this particular issue.
[13] See also Ali and Seiford (1993) for another second stage LP method for the measurement of the slacks.
[14] For a general discussion of this problem see, among others, Coelli *et al.* (1998) and Fried *et al.* (1993).
[15] Charnes, Cooper, Golany, Seiford and Stutz (1985). See Charnes, Cooper, Rousseau and Semple (1988) for a unit invariant version of the additive model.

Another alternative to the 2-stage process of the orientated models is the "multistage DEA" proposed by Coelli (1998), which conducts a sequence of radial movements to identify the efficient peer of the unit under observation. This methodology has the double advantage of being invariant to units of measurement as well as preserving as much as possible the original input and output mixes, which makes the suggested changes for the unit under observation more realistic.

The assumptions of convexity of the feasible set, as well as that of strong disposability of inputs and outputs are very rarely relaxed[16], whereas the possibility of having log-linear envelopment surfaces is performed by the "multiplicative" models as for example in Charnes *et al.* (1982, 1983). Finally, more recent contributions to the implementation of a stochastic DEA can be found in the literature[17].

[16] The convexity assumption is relaxed in Deprins *et. al.* (1984) and Tulkens *et al.* (1990), where the convex hull is replaced by a free disposable hull.

3.3 The econometric estimation of frontiers.

The econometric estimation of frontiers[18] is an approach that was developed out of a criticism of the econometric estimation of production functions[19], well established since the early work of Cobb and Douglas. The estimation of a production function is based on the assumption that all producers in the sample are behaving efficiently, and that any deviation from the regression line is due only to statistical noise. In other words, what one estimates is an "average" production function in which the parameters representing the technique are by definition the same for all observations in the sample. The production frontier approach stresses the point that this might, and most probably would, not be the case. This new parametric approach then developed into different specifications, whose common characteristic is that the inefficiency component is modelled by an error term.

Even though the object of this chapter are the stochastic frontier models, for completeness and ease of explanation an overview of the deterministic frontiers will be given first.

In the deterministic frontier case, the general set up of the problem is to estimate

$$y = f(x;\beta)D_o \tag{3.9}$$

where y is the level of output, x is a set of inputs and β a vector of parameters to estimate; $f(x;\beta)$ is assumed to be smooth, continuous, continuously differentiable

[17] For example an application to the electricity distribution is in Weyman-Jones (mimeo).
[18] Comprehensive reviews of the topic can be found in Greene (1997) and in Kumbhakar and Lovell (2000).
[19] Obviously, the same goes for cost, or profit, or revenue functions.

and quasi concave[20]. Finally, D_o is inefficiency: as in (3.6) inefficiency is measured by the output distance function, defined as the ratio of actual to efficient output, that is $y/f(x;\beta)=D_o\leq 1$

Empirically, this usually translates into the estimation of a log-linear function, which in the single-equation, cross-sectional case is

$$lny_i = \alpha + \beta' lnx_i - u_i \qquad (3.10)$$

$i = 1,\ldots,N$

where N is the total number of observations, β is a Kx1 vector of parameters to estimate, y and x are defined as in (3.9) and $u_i = -lnD_o$.

Inefficiency is measured by the vector of i.i.d. random variables u_i, independent of the regressors. From (3.9) and (3.10) it is clear that $u_i \geq 0$. In fact, using the definition of distance function and the notation from (3.6) one can rewrite (3.10) as

$$lny_i = \alpha + \beta' lnx_i + lnD_{oi} \qquad i=1,\ldots,N \qquad (3.11)$$

where

$0<D_0\leq 1$ is the distance from the (output) frontier

$D_o = e^{-u}$

and therefore

$lnD_o = -u$ or $-lnD_o = u$

In other words, this error component must come from a non-negative distribution.

[20] I.e. the properties of a well behaved production function are assumed.

This deterministic model was first proposed by Aigner and Chu (1968), and several contributions to the estimation of (3.11) followed their paper. One possibility is for example to adjust the results of an estimation carried out by ordinary least squares (OLS). OLS gives consistent estimates of all the parameters but the intercept. Gabrielsen (1975) and Richmond (1974) proposed two adjustments to the OLS estimation, respectively the "corrected least squares" (COLS) and the "modified least squares" (MOLS), both consisting of a shift of the regression line upwards on the basis of the calculated residuals. COLS adjusts the regression line upwards until the largest residual is 0. That is, the intercept in (3.10) and the OLS residuals are adjusted as

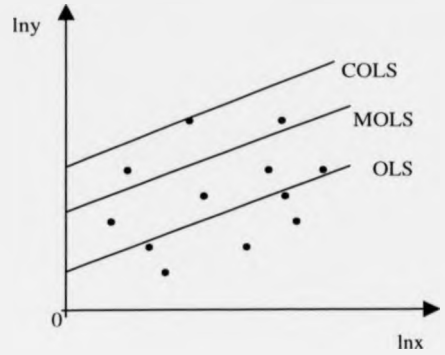
$$\begin{aligned} \hat{\alpha}^* &= \hat{\alpha} + \max_i[\hat{u}_i] \\ -\hat{u}_i^* &= \hat{u}_i - \max_i[\hat{u}_i] \end{aligned} \tag{3.12}$$

MOLS makes an assumption about the distribution of u_i and the adjustment is based on its estimated mean, extracted from the moments of the OLS residuals. Again, in terms of equation (3.10) this means that

$$\begin{aligned} \hat{\alpha}^* &= \hat{\alpha} + E[\hat{u}_i] \\ -\hat{u}_i^* &= \hat{u}_i - E[\hat{u}_i] \end{aligned} \tag{3.13}$$

As a consequence, in MOLS some of the residuals might still be positive, i.e. some observations lie above the frontier (see Fig.3.6).

Fig.3.6: Example of regressions performed by OLS, MOLS and COLS, with one input x and one output y (in logs).



In both cases the frontier has the same parameters as the OLS regression line with the only exception of the intercept. This means that all producers, whether efficient or inefficient, are considered to have the very same technology; the ranking of units is the same as that of OLS, and the only difference is the absolute value of the distance from the frontier (the value of the adjusted residuals).

This implication can be too strong, and the problem can be overcome if (3.11) is estimated by maximum likelihood (ML). In this case a distribution function for u_i has to be assumed to respect the non-negativity requirement, and it is therefore taken into account when the parameters are estimated. The possibility of using ML was first proposed by Afriat in 1972. Various distributions were subsequently discussed by others, like the half-normal and the exponential (Schmidt, 1976), the truncated normal (Stevenson, 1980) and the gamma (Greene, 1990). The characteristics of different possible distributions will be discussed shortly, as they are common to the stochastic frontier models.

Apart from specific issues related to the distributional assumptions of the inefficiency component, all the deterministic models suffer from their very nature, as all deviations from the frontier are attributed to inefficiency, no account being taken of statistical noise, measurement errors or factors not under the control of the firm. In a way, this is the very opposite problem of estimating a production function: in one case all deviations are attributed only to noise, in the other they are attributed only to inefficiency. This limitation opened the way to the formulation and estimation of stochastic frontiers.

The stochastic frontier model was proposed independently and at the same time by Aigner, Lovell and Schmidt (1977), Battese and Corra (1977) and Meeusen and Van den Broek (1977). The idea is to introduce in the equation a stochastic component together with the inefficiency component. This translates into a composite error term given by their sum. Again in the single equation, cross sectional case this is

$$lny_i = \alpha + \beta' \, lnx_i + \varepsilon_i \qquad (3.14)$$

where x, y, i and β are defined as in (3.10); the composite error term is

$$\varepsilon_i = (v_i - u_i)$$

where

$$v_i \sim N(0, \sigma_v^2)$$

is the stochastic component, a vector of independently and identically distributed normal random variables, with a zero-mean and constant variance σ_v^2. Inefficiency is measured by the vector of random variables u_i assumed to be independent of the v_is and of the regressors.

$(\alpha+\beta' \ lnx_i)$ is now the deterministic part of the equation, as in deterministic models $v=0$, and $(\alpha+\beta' \ lnx_i+v_i)$ is the stochastic frontier.

The composite disturbance ε_i resulting from the sum of statistical noise and inefficiency is asymmetrically distributed with a negative skew, its final distribution depending on the distribution assumed for u_i. Because of the presence of a composite error term the use of OLS would give consistent but inefficient parameters' estimates, as well as a biased intercept. This last problem could be overcome by using techniques like MOLS, but the problem of inefficiency remains, and for this reason, if the distribution of u_i is known (or rather an assumption is made about it), ML estimation is to be preferred[21].

As was anticipated for the deterministic frontiers, the most common distributions that are found in the literature are the half normal, the truncated normal and the exponential[22].

In the case of the half normal distribution (Schmidt, 1976), u_i is assumed to be the absolute value of a normally distributed variable with a zero mean, and then

$\varepsilon_i = v_i - u_i$

$v_i \sim N(0, \sigma_v^2)$

$u_i = |U_i|$

$U_i \sim N(0, \sigma_u^2)$

[21] As of the finite sample properties of the two estimation methods, so far a Monte Carlo simulation by Coelli (1995) is what can be found in the literature to prove the superiority of ML compared to MOLS.

[22] A gamma distribution was first attempted by Greene (1980), and it showed computational difficulties. For more details on the literature about it see for example Kumbhakar and Lovell

The log-likelihood function of the frontier model, with its composite error term, is (Aigner, Lovell and Schmidt, 1977)

$$L = -N \ln \sigma - A + \sum_{i=1}^{N} \left\{ \ln \Phi \left[\frac{-\varepsilon_i \lambda}{\sigma} \right] - \frac{1}{2} \left[\frac{\varepsilon_i}{\sigma} \right]^2 \right\} \tag{3.15}$$

where

$$\varepsilon_i = \alpha + \beta' \ln x_i - \ln y_i$$

$$\lambda = \frac{\sigma_u}{\sigma_v}$$

$$\sigma^2 = \sigma_u^2 + \sigma_v^2$$

A is a constant and Φ is the distribution function of the standard normal. The parameter λ[23] embodies the influence of the inefficiency component, but as will be seen later a different, computationally more convenient parameterisation can be used, as in Battese and Corra (1977).

As observed by Stevenson (1980), a half normal distribution is equivalent to the truncation at 0 of a normal variable with a 0 mean. The case could be generalised to the truncation at zero of normal distributions with a non-zero mean, because the zero-mean could be an unnecessary restriction. The resulting truncated normal distribution depends therefore on whether the mean is positive or negative. More in detail for this case

$u_i = |U_i|$

and

(2000). Some progress has been very recently made in Greene (2000) by using a simulated maximum likelihood estimation, as opposed to the direct maximisation.

$U_i \sim N(\mu, \sigma_u^2)$

and the log-likelihood of the frontier equation is

$$L = -N\left[\ln\sigma + \frac{1}{2}\ln\pi + \ln\Phi\left(\frac{-\mu}{\sigma\lambda}\right)\right] - \sum_{i=1}^{N}\left\{\frac{1}{2}\left[\frac{\varepsilon_i}{\sigma}\right]^2 - \ln\Phi\left[\frac{-\mu}{\sigma\lambda} - \frac{\varepsilon_i\lambda}{\sigma}\right]\right\} \qquad (3.16)$$

where σ, λ and Φ are defined as in (3.15).

Finally, another distribution for u_i that still maintains the non-negativity requirement is the exponential one, as in Aigner et al (1977) and Meeusen and van den Broek (1977). In this case

$$h(u_i) = h\exp(-hu_i)$$
$$h, u_i > 0$$

and the corresponding log-likelihood is

$$L = N\left[\ln h + \frac{1}{2}(h\sigma_v)^2\right] + \sum_{i=1}^{N}\left[\ln\Phi\left(\frac{-\varepsilon_i}{\sigma_v} - h\sigma_v\right) + h\varepsilon_i\right] \qquad (3.17)$$

The choice of the distribution of u_i obviously affects the estimation of inefficiency. In particular the half normal and the exponential distributions both have the mode at 0, which means that the probability is highest of having inefficiency effects equal to 0, i.e. to estimate firms as efficient. A more general distribution as the truncated normal doesn't suffer from this problem. As the half and truncated normal distributions are nested models, with the former being equivalent to the latter having a 0 mean, the null hypothesis H_0: $\mu = 0$ can be tested against the alternative hypothesis H_1: $\mu \neq 0$ by means of a Likelihood

[21] The same notation as in the original paper has been used here; the parameter λ has nothing to do with that of (3.7) and (3.8).

Ratio (LR) test, and so a choice can be made between the two models on statistical grounds. As is known, the LR test is specified as

$$LR = 2[\mathcal{L}(H_1)-\mathcal{L}(H_0)]\sim\chi^2_r \qquad (3.18)$$

where $\mathcal{L}$ is the value of the maximised log-likelihood. This follows a χ^2_r distribution with r degrees of freedom, where r is the number of restrictions (in this case $r = 1$).

This testing procedure cannot be used to compare the exponential distribution and the (truncated or half) normal one, because the models are non-nested. In this case, information criteria for choosing between non-nested models can be used, like for example the Akaike information criterion. This is based on the comparison of the values of the maximised log-likelihood functions, taking into consideration the number of parameters of each specification, in order to consider also how parsimonious a model is. The Akaike information criterion is specified as

$$AIC = -2\mathcal{L} + 2n \qquad (3.19)$$

Again, $\mathcal{L}$ is the value of the maximised log-likelihood and n is the number of parameters. The preferred model is the one with the lowest AIC value.

The possibility of making a choice between different distributions should obviously be welcomed, and the fact that this is not always a proper statistical test is a shortcoming. However, there is some evidence (Greene 1990; Kumbhakar and Lovell, 2000) that the rankings of producers are not particularly sensitive to the choice made.

For estimation purposes[24] a useful parameterisation to measure the influence of the inefficiency component is that proposed by Battese and Corra (1977). This is done via the definition of a parameter $\gamma = \sigma^2_u/\sigma^2$, instead of the λ parameter used in (3.15). The corresponding log-likelihood for the case of a half normal distribution of u_i is

$$L = -\frac{N}{2}\left(\ln 2\pi + \ln \sigma^2\right) + \sum_{i=1}^{N} \ln \Phi\left[-\left(\frac{\varepsilon_i}{\sigma}\right)\sqrt{\frac{\gamma}{1-\gamma}}\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i^2 \qquad (3.20)$$

$$\gamma = \frac{\sigma_u^2}{\sigma^2}$$

$$\sigma^2 = \sigma_u^2 + \sigma_v^2$$

and therefore

$\gamma \in [0,1]$

A value of γ equal to 0 means that all deviations from the frontier are due to noise, whereas a value equal to 1 means that they are due only to inefficiency and the frontier is actually deterministic. Testing for the significance of inefficiency is therefore testing for the null hypothesis H_0: $\gamma = 0$ against the alternative hypothesis H_1: $\gamma \neq 0$. This can be done by means of an LR test. In this particular case, however, as the 0 value lies on the boundary of the parameter's space the statistic follows a mixed χ^2_r distribution, and its critical value for a test of size (α) corresponds to that of a test of size (2α).

As regards the measurement of each firm's inefficiency, which is the purpose of the whole exercise, the problem arises that what one is interested is u_i, (and

every firm's inefficiency will then be exp[$\hat{u}_i$]) but the residuals of the regression are instead $\hat{\varepsilon}_i$. So far, the best proposed solution to this problem is to calculate the conditional probability $E[u_i|\varepsilon_i]$; this gives an unbiased though inconsistent measure of u_i, because the variance of the estimate remains non-zero as it is independent of N[25]. A better estimate can be obtained in the case of panel data models, which will be discussed next.

3.3.1 Panel data.

The availability of a panel data set has some desirable properties for the estimation of a stochastic frontier. Three main advantages in particular are pointed out by Schmidt and Sickles (1984), all consequent to having several observations on the same cross sectional unit: no distributional assumptions on the composite error term are necessary anymore, there is no need to assume that inefficiency is uncorrelated with the regressors (though this is true only in the fixed effects model) and finally inefficiency can be estimated consistently if $T \rightarrow \infty$ (a benefit though that only long panel data sets show, which is not often the case).

The presence of a time dimension opens up different possibilities for the estimation of both the parameters and the inefficiency component itself. The literature on the various possible models is quite large on its own, but for the

[24] Greene (1997).
[25] For details on the possible estimators proposed by the literature see for example Kumbhakar and Lovell (2000).

purpose of this chapter the discussion will be kept fairly general, to provide an overview of the advantages and characteristics of different specifications.

In the presence of panel data, assuming parameters remain constant across cross sectional units and over time, the model to be estimated is

$$lny_{it} = \alpha + \beta' lnx_{it} + \varepsilon_{it} \tag{3.21}$$

$i = 1,....,N$

$t = 1,...,T$

that is the N firms are each observed T times (for a balanced panel, but the analysis can be extended to unbalanced panels as well). The structure of ε_{it} depends on the assumption made about inefficiency. If inefficiency is time-invariant then

$$\varepsilon_{it} = v_{it} - u_i \tag{3.22}$$

whereas if it is time-varying then

$$\varepsilon_{it} = v_{it} - u_{it} \tag{3.23}$$

In both cases the stochastic component is v_{it}, assumed to be

$v_{it} \sim N(0, \sigma^2_v)$

Starting with the case of a time-invariant inefficiency, like (3.22), the usual fixed effects (FE) or random effects (RE) models can be used. In the FE model, u_i is assumed to be a fixed, producer-specific constant. No distributional assumption is therefore necessary about u_i, and this can be correlated with the regressors and/or with the stochastic component, which are both advantages over the ML estimations discussed before for the cross sectional case.

For a FE model, (3.21) can be estimated by least squares with dummy variables (LSDV), with an adjustment similar to that used in COLS to maintain the non-negativity constraint on u_i. In particular, equation (3.21) is rewritten as

$$lny_{it} = \alpha_i + \beta' lnx_{it} + v_{it} \tag{3.24}$$

where $\alpha_i = (\alpha - u_i)$ are the firm-specific intercepts. Once the model has been estimated the adjustment to ensure that the non-negativity constraint on the inefficiency component holds is

$$\hat{\alpha} = \max[\hat{\alpha}_i]$$
$$\text{and} \tag{3.25}$$
$$\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i$$

LSDV gives consistent parameters estimators as either N $\rightarrow \infty$ or T$\rightarrow \infty$, and consistent estimators of u_i if both N$\rightarrow \infty$ and T$\rightarrow \infty$. This is an advantage over ML , although in practice T is often quite short.

The drawback of this estimation of FE is that the constant, producer-specific u_i can capture any other producer-specific, time-invariant characteristic, wrongly attributing it to inefficiency, but no time invariant regressors can be included in the equation.

If a RE model is used, u_i is assumed to be a random variable rather than a fixed constant. In this case the ordinary least squares estimators are inefficient, and the model is better estimated by generalised least squares (GLS). This in turn makes it necessary to assume non-correlation between the u_is and both the regressors and the stochastic component, although time-invariant regressors can be included in the equation, avoiding the problem seen for the FE model.

For the RE model the equation to estimate is

$$lny_{it} = \alpha^* + \beta' lnx_{it} + v_{it} - u_i^* \tag{3.26}$$

where

$$\alpha^* = \{\alpha - E[u_i]\} \tag{3.27}$$

and

$$u_i^* = \{u_i - E[u_i]\} \tag{3.28}$$

Once all parameters in (3.26) have been estimated by GLS the u_i^* can be estimated from the residuals and finally

$$\hat{u}_i = \{\max[\hat{u}_i^*] - \hat{u}_i^*\} \tag{3.29}$$

Again, this provides consistent estimators of u_i as both $N \rightarrow \infty$ and $T \rightarrow \infty$.

Alternatively, if distributional assumptions can be made as well as the non-correlation one, ML provides overall more efficient estimators than both FE or RE models, and consistent estimators of u_i as $T \rightarrow \infty$. The distributional assumptions about u_i are the same discussed for the cross sectional models seen before, so no further detail is needed here.

It appears from the above that there are different advantages and disadvantages in using the three techniques, and there are no a priori, general reasons to prefer one to the others in absolute terms. The choice will depend on the nature of the data and on the assumptions that in every case it is reasonable to make.

These main considerations also hold when inefficiency is time-varying. In this case, the main difference is that a functional specification to represent the variation over time has to be made, which is among the reasons whether a particular model will be best estimated by LSDV or by GLS or again by ML. No further detail is provided here about the different models found in the literature, with the exception of the model presented by Battese and Coelli (1992), which is the one used in Chapter 5[26]. The functional form chosen for the inefficiency component in their paper is given by

$$u_{it} = u_i \exp[-\eta(t-T)] \tag{3.30}$$

where u_i comes from a non-negative distribution (the model is estimated by ML) and T is the last year of the observations. This formulation expresses the inefficiency of each firm as a function of its value at time T. From (3.30) when $t = T$ then $u_{it} = u_i$. A value of $\eta>0$ means that u_{it} decreases over time, i.e. efficiency increases, and viceversa when $\eta< 0$. A value of $\eta = 0$ means there is no time effect, and the model reduces to a RE model with $u_{it} = u_i$. The hypothesis can be tested by means of an LR test.

This time-varying specification can be automatically performed by the software FRONTIER 4.1[27], used in Chapter 5.

Finally, a possible extension to the models presented so far is worth mentioning, which is the issue of trying not only to measure but also to explain differences in (in)efficiency. The first attempts to do that made use of a two-stage analysis, where the second stage would consist of a separate regression of the estimated

[26] For a discussion of other panel data models, see Kumbhakar and Lovell (2000) and the references therein.
[27] Coelli, T. J. (1996a), "A guide to FRONTIER version 4.1".

(in)efficiencies on a vector of exogenous variables. This two-stage approach showed various econometric problems, so alternatives were proposed in which the inefficiency component would be explained by a set of exogenous variables estimated at the same time as the other parameters by ML. The work of Deprins and Simar (1989) was among the first in this area, with the estimation of a deterministic frontier. Others followed, like for example Kumbhakar, Ghosh and McGuckin (1991) with an application to the stochastic frontier context, and Battese and Coelli (1995) with a panel data set.

3.4 Technological change, shifts of the frontier and the Malmquist index.

When the efficiency of a set of firms is assessed over several periods of time, the question arises quite naturally of whether time itself might have had an effect on their behaviour. This is more so the longer is the time period, or if some relevant, external events took place during it. If no technological change had occurred at all, then one could estimate one frontier only and all recorded changes in the performance of the firms could be interpreted as changes in technical efficiency only. On the other hand, if this is not true, the frontier itself might have changed (usually shifted) over time, and separate estimations for every year are appropriate. It is interesting in this case to separate the changes in the frontier itself, i.e. technological change, from the changes in technical efficiency of every unit observed.

A change in total factor productivity (TFP) is a change in the ratio of outputs to inputs. The use of index numbers to measure this change dates back to the works

of Fisher (1922) and Törnqvist (1936), that defined ratios between the (indexes of) outputs and the (indexes of) inputs of a firm at different points in time. Since the works of Caves, Christensen and Diewert (1982a, 1982b), that decomposed it into various components, a very common way of measuring TFP change is by means of a Malmquist index (Malmquist, 1953). This is calculated as the ratio of distance functions, and so it assumes their estimation beforehand. This methodology has become very popular when efficiency is calculated by DEA[28].

The Malmquist index is a summary measure of the change in TFP of a given unit over time. This overall measure can be split up as the product of three different components: the change in technical efficiency (measuring whether the unit has moved closer to the frontier), the change in scale efficiency (measuring whether the unit has moved closer to the constant returns to scale facet of the frontier) and the shift of the frontier itself (measuring whether the unit has improved its production possibilities). This is calculated as ratios of distance functions. In more detail, let's recall the definition of output distance function given in (3.6)

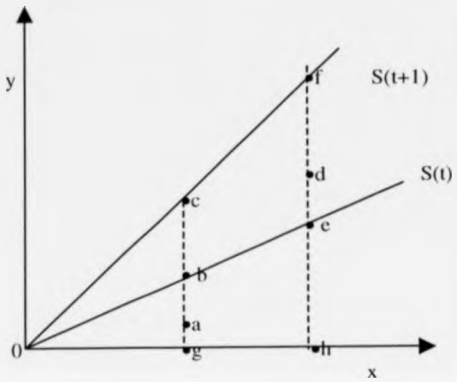
$$D_o = \min\left\{\vartheta : \frac{y}{\vartheta} \in P(x)\right\}$$

A value of $D_o=1$ implies that a unit is located on the frontier, and a value <1 that it is below it. Let's now assume that a given unit i that uses inputs vector x to produce output vector y, is observed over two different times, t and $t+1$. For simplicity the situation is represented for the one-input and one-output case, with

[28] Some applications to econometric frontiers are reviewed in Grosskopf (1993) and in Coelli *et al.* (1998). However, the model estimated in Chapter 5 does not allow for it.

constant returns to scale in Fig. 3.7, where *S(t)* and *S(t+1)* are the frontiers at time *t* and *t+1* respectively.

Fig.3.7: Change of the frontier S (one output and one input case) between time t and t+1.



Unit *i* at time *t* is point *a* in the figure and at time *t+1* it is point *d*.

Let's now define the following:

$D_{ci}^{t}(x_i^t, y_i^t)$ is the distance of the unit as observed at time *t* (i.e. using input vector x_i^t to produce output vector y_i^t) from the frontier of time *t*, *S(t)*. This corresponds in the figure to the segment *ab*, in turn the ratio *ga/gb*.

$D_{ci}^{t+1}(x_i^{t+1}, y_i^{t+1})$ is the distance of the unit at time *t+1* from the frontier of time *t+1*, *S(t+1)*. This corresponds to the segment *df* (ratio *hd/hf*)

$D_{ci}^{t+1}(x_i^t, y_i^t)$ is the distance of the unit observed at time *t* from the frontier of time *t+1*: segment *ac* (ratio *ga/gc*)

$D_{ci}^{t}(x_i^{t+1}, y_i^{t+1})$ is the distance of the unit observed at time t+1 from the frontier of time *t*, corresponding to the segment *ed* (ratio *hd/he*).

Given the above definitions, the Malmquist index for the particular unit considered is defined as

$$M_{oi} = \frac{D_{oi}^{t+1}(x_i^{t+1}, y_i^{t+1})}{D_{oi}^{t}(x_i^{t}, y_i^{t})} \left[\frac{D_{oi}^{t}(x_i^{t+1}, y_i^{t+1})}{D_{oi}^{t+1}(x_i^{t+1}, y_i^{t+1})} \frac{D_{oi}^{t}(x_i^{t}, y_i^{t})}{D_{oi}^{t+1}(x_i^{t}, y_i^{t})} \right]^{\frac{1}{2}} \tag{3.31}$$

which in terms of Fig. 3.7 is equivalent to

$$\frac{\left(\frac{hd}{hf}\right)}{\left(\frac{ga}{gb}\right)} \left[\frac{\left(\frac{hd}{he}\right)}{\left(\frac{hd}{hf}\right)} \frac{\left(\frac{ga}{gb}\right)}{\left(\frac{ga}{gc}\right)} \right]^{1/2}$$

The term outside the brackets represents the change in technical efficiency, that is whether the unit has moved closer to its frontier. The part inside the brackets represents technical progress, calculated as the geometric average of the distance from the two frontiers. If variable returns to scale are assumed, the part outside the brackets can be split in turn into a change in the scale of operation and a real change in efficiency.

A Malmquist index bigger than 1 indicates that total factor productivity has increased, and viceversa for a value smaller than 1. The specific changes can be interpreted in a similar fashion, with ratios of distance functions bigger than 1 indicating an improvement and viceversa.

As regards econometric estimations, technological change is usually modelled by introducing a time effect among the regressors[29]. This can be in the form of a time trend or of dummy variables to allow for a different intercept for different years. In both cases the slope parameters of the regression equation are assumed to be constant, over time and across cross sectional units. The possibility however that the very shape of the production function might at some point have changed is an interesting one. This possibility is explored in Chapter 5, where a structural break is detected by means of a time interaction-dummy. This dummy takes a value of 1 for particular years and 0 else, and it is multiplied to all the variables (the $x_{it}s$) in the regression equation, i.e.

$$lny_{it} = \beta' \, lnx_{it} + d + \rho' \, lnx_{it} \, d + \varepsilon_{it}$$

where β and ρ are two Kx1 vectors of parameters to estimate and d is the time interaction dummy. This means that the parameters of the function will be the vector $\beta+\rho$ when $d = 1$, and the vector β when $d = 0$. As the details on this cannot be separated from the analysis, to avoid repetition the issue is postponed to Chapter 5.

3.5 The efficiency of hospital services.

This last section will highlight the main issues and literature contributions to the estimation of efficiency in the hospital sector.

The estimation of efficiency in the hospital sector raises some peculiar problems especially as regards the definition and measurement of output (see for example

[29]Comprehensive reviews of different models are in Grosskopf (1993) and Coelli *et al.* (1998).

McGuire, 1985). First of all, it is not possible to consistently measure the final output of hospitals, the improvement in health, so that an intermediate measure of it is required. This not only disregards the final aim of hospitals activity; isolating hospitals and their output from the rest of the health care system, it also fails to give any consideration to the linkages and interrelations within it, which are relevant to the degree of integration. The more integrated is the system, the less clear are the boundaries between all the services and the bigger the possibility of spreading health care among them, and this, in turn, makes it more complicated to talk about efficiency (Evans, 1981).

The main problem with the intermediate output is that it is not homogeneous, but varies widely across hospitals and even within each of them. Two main approaches to the homogenisation of output can be found in the literature (Tatchell, 1983). The service-mix approach measures output in terms of the services actually or potentially provided by the hospital, thus focussing on the inputs to identify the outputs. The case-mix approach, much more widely used, identifies the output in the number and (diagnostic) kind of cases treated by the hospital. These translate into the definition of different casemix categories, even though heterogeneity remains on other factors, like the severity of illness within each category. However, too detailed a level of output definition has the shortcoming of increasing greatly the number of parameters to estimate, so that a trade-off between precision and statistical efficiency is usually unavoidable (Butler, 1995).

The presence of multiple outputs is not a problem for DEA, but it can be so in econometrics, as the production frontier models analysed before can be estimated only for the single-output case.

Three solutions can be found in the literature to overcome this problem

1) The estimation of a cost, as opposed to a production, frontier.

2) The use of index numbers.

3) The estimation of a distance function.

The first solution has been frequently applied in the literature. The estimation of a cost function has one variable only on the LHS, the total cost of production, and all outputs together with the input prices are on the RHS. When estimated as such, the measured (in)efficiencies will represent both technical and allocative (in)efficiency, but the two can be disentangled by estimating simultaneously the cost frontier and the factors share equations. There are two problems with this approach: the very high number of regressors (especially if one uses a flexible functional form) and the definition of input prices.

The second solution is very appealing when one is not interested in the marginal effects of the different outputs per se. The main problem associated with this approach is that of finding suitable weights for the construction of the index.

The third solution has appeared more recently in the literature, and it is the one chosen for the analysis in Chapter 5[30], where in particular the model proposed by

[30] Some use of index numbers will prove necessary too.

Coelli and Perelman (1996) is used[31]. The discussion of the model is done in that chapter, together with some considerations related to the choice of a functional form for the frontier.

Coming to the empirical literature on hospitals' efficiency, this has focussed in particular on the analysis of their costs. A more traditional approach, that starts with Feldstein's seminal work (Feldstein, 1967), estimated what Evans defined "behavioural" cost functions (Evans, 1971), as opposed to the more recent estimation of "proper" cost functions (i.e. resulting from a constrained minimisation problem). The works using this second approach, mostly from the USA, are estimations of cost functions (Cave *et al.*, 1978; Wagstaff, 1989), often estimated simultaneously with the factor share equations to disentangle technical and allocative efficiency (Cowing and Holtmann, 1983; Conrad and Strauss, 1983; Fournier and Mitchell, 1989). The estimation of *frontiers* is more recent, both for DEA and the stochastic frontier approach.

Given the object of this thesis, it is worth focussing attention on the analyses of the UK hospital sector.

The literature on the efficiency of the UK hospital sector was surprisingly not very rich before the reform (Wagstaff, 1988). McGuire and Westoby (1983) estimated a translog production function on Scottish, non-teaching acute hospitals, with a focus on the efficiency of the input-mix which showed the existence of an excess in capital expenditure and housekeeping services. Gray *et*

[31] One application of this model is in Burns *et al.* (2000).

al. (1986) analysed trends in factor inputs in Scottish hospitals between 1951 and 1981, revealing a shift towards cheaper labour factor inputs.

An upsurge of interest on the topic was marked by the introduction of the reform. Soderlund *et al.* (1997) used a linear regression model (not a frontier) on a sample of NHS hospitals in England for 1992-1994, which revealed a general productivity improvement whose association with the changes to trust status remained however unsure.

A lot of the literature used the same Scottish data set as in the present work. For example Scott and Parkin (1995) used it for 1992/93 to estimate a translog cost function which highlighted the prevalence of constant returns to scale and economies of scope between different kinds of outputs (mainly inpatients and outpatients). Parkin and Hollingsworth used DEA on this data set for the period 1991/92 to 1993/94. Their work analysed the strengths and weaknesses of DEA, especially as regards its sensitivity to different aggregation methods for inputs and outputs. No trends or indexes of productivity change were calculated.

The most similar work to the present one, to our knowledge, is Maniadakis *et al.* (1999) which used the same data set for the period 1992-1996[32]. Their analysis consists of the calculation of DEA-based Malmquist indexes of TFP, on a sample of 75 acute hospitals in Scotland. The paper also allows for a measure of quality[33], and concludes for a worsening of it over time. Their general results are similar to the ones of this thesis, as regards the overall improvement in TFP

[32] Following the analysis started by Maniadakis and Thanassoulis (1997).
[33] In particular, it is the survival rate after 30 days from discharge.

which is mainly attributable to shifts of the frontier. Their calculated improvement is larger (+7%) probably because of the different output categorisation used (more emphasis is given to the outpatients, with only one category of inpatients) and for other statistical reasons discussed below. No separate analysis of the role of trusts is made, nor of the role of different inputs in explaining inefficiency. This translates into more optimistic conclusions about the effectiveness of the reform than the ones drawn in this thesis.

The differences and contributions of the present work are therefore worth mentioning here.

First of all, the sample size is different. The time span covers the whole duration of the reform. As regards the number of cross sections, the original total number (that is the total number of hospitals classified as acute) was 75, and this larger sample was used in the papers mentioned above. However, a closer look at the data disclosed a problem. Many of these hospitals, though registered as acute, are actually cottage hospitals; these are a very particular category as they are very small, do not perform any kind of surgery and have different staffing procedures, as they often rely upon local GPs. It therefore seemed rather inappropriate to compare them with larger, general hospitals, especially when using a deterministic technique. Moreover, when looking at the data series for this subset of the sample, they showed lots of irregularities and inconsistencies. The matter was further investigated by contacting the hospitals directly and it was realised that the data were very inaccurate for this part of the sample. For these reasons it was finally decided to remove these observations: even if this would have a cost

in terms of degrees of freedom, it would obviously give more reliable and meaningful results.

Secondly, and maybe most importantly, the kind of statistical analysis performed is different. In the above non parametric papers, all general conclusions about trends in changes were drawn from the analysis of the average of the indexes results. As will be seen in Chapter 4, this can be inappropriate (and turned out to be so) if the distribution of the results themselves is not checked. DEA is a deterministic technique (so no noise is considered) and is very sensitive to the presence to outliers. If such outliers are not checked for and taken into account, the results can be biased and the conclusions therefore misleading.

Finally, also the stochastic, parametric approach is used, as the two techniques are quite complementary in terms of strengths and weaknesses. This is thought to make the analysis more reliable in terms of general results, and more complete as it opens up different possibilities. For example, the changes in the relative inputs inefficiencies and/or elasticities are calculated, the adoption of a different technology of production is analysed, none of which has been done before.

The next two chapters are devoted to the actual estimation of efficiency, using respectively DEA in Chapter 4 and the SF in Chapter 5. The comparison of the two methodologies is done in Chapter 6, together with the comparison of the results and the conclusions.

CHAPTER 4

DEA AND MALMQUIST INDEXES ANALYSIS

This chapter performs an analysis of the changes in productive efficiency of a sample of 53 acute hospitals in Scotland in the period 1991/92 - 1996/97; the aim is to see whether the reform of the NHS, which in 1990 introduced competition for hospital services, has actually improved the efficiency with which hospitals perform their activity. The data do not cover the years preceding the introduction of the reform; however, the acquisition of trust status by hospitals, which embodies the full working of the internal market, did not take place all at once, so that the sample contains both trusts and non-trusts which can be compared.

The methodology used is the calculation of Malmquist indexes of total factor productivity change, based on non-parametric frontiers estimated by Data Envelopment Analysis.

Methodological issues have already been discussed in Chapter 3. This chapter therefore starts with the description of the data set, in Section 4.1. The overall change between 1991/92 and 1996/97 is the subject of Sections 4.2 and 4.3, devoted respectively to the analysis of the Malmquist index and the determinants of inefficiency. Year by year changes are analysed and discussed in Section 4.4 and general conclusions are provided in Section 4.5.

4.1 The data.

The data are a sample of 53 acute hospitals in Scotland in the years 1991/92 (from now on referred to as 1992) to 1996/97 (from now referred to as 1997), that make a panel data set of 6x53 = 318 observations.
These data were obtained from the Scottish Health Service Costs statistics.

As regards the definition of the inputs and outputs variables, the following choices have been made (Tatchell, 1983; Butler, 1995).

Output is measured as the total number of cases treated in various specialty (or casemix) categories. The original data set is quite detailed, with many different categories and a main distinction between the patients who need to stay in the hospital overnight (inpatients), and those who do not (outpatients, day patients and day cases). This main distinction has been kept. The many categories of inpatients have been in turn summarised into three general groups in order to reflect at least some of the difference in the casemix of different hospitals. This translated into 4 output categories:

q_1= inpatients, surgery (total number of cases);

q_2= inpatients, medical (total number of cases);

q_3=inpatients, others (total number of cases);

q_4= outpatients, day cases and day patients (total number of cases).

The inputs are defined by the following 5 variables

x_1=Total capital charges (£000);

x_2= Medical staff WTE (whole time equivalent);

x_3= Nursing staff WTE;

x_4= Other staff WTE;

x_5= Total number of beds.

The measure on capital comprises:

a) Depreciation on fixed assets;

b) Interest paid on money borrowed to finance any of the projects in a).

c) 6% return on capital (trusts only).

Capital is measured in £000, and it is deflated using the "Hospital and Community Health Services pay and price inflation values".

The "other staff" input includes professional, technical, administrative, clerical and all other staff.

The "whole time equivalent" is the number of staff expressed in relation to the standard weekly hours for a particular staff category.

Tables 4.1 and 4.2 report the summary statistics for the sample, respectively for the outputs and the inputs. The distribution of the data series was checked and they showed to be extremely asymmetric, with a very wide variation around the mean[1]. This can translate in the average being biased by the presence of some outliers and in turn be inaccurate to represent changes in the levels over time. As

[1] For reasons of space the graphs of the distributions are not presented.

a consequence, an adjusted average measure is used instead, which includes only the observations within $\mu \pm 2\sigma$ (where μ and σ are the mean and the standard deviation of the distribution)[2]. For the same reason and for completeness also the median is reported.

As can be seen in Table 4.1, the average number of patients treated increased in three out of the four categories. In particular, the inpatients in the categories surgery and medical show quite a regular and steady increase, and the same goes for the outpatients and day patients, whereas the category "inpatients others", i.e. all other kinds of treatment provided by the hospitals that require the patient to stay overnight, presents a negative trend.

This could be in line with the general expectation that hospitals might have tried to reduce the number of inpatients and/or the length of stay as much as possible for cost saving reasons, and resort to provide treatment outside the hospital or without the need of overnight stay (that is the need of a staffed bed). It can be guessed that this is easier on the category "others" which probably includes less serious kinds of illness.

As regards the inputs values, Table 4.2 shows that over time the levels of capital, medical staff and other staff increased, whereas the number of beds and of nursing staff decreased.

[2] According to Tchebichev's rule this is at least 75% of the population, a value that raises to 95% if the distribution is normal, or close to normal.

The acquisition of trust status brought with it some accounting changes as regards the capital input, so part of this increase can be directly related to that. The pattern of change in the levels of this variable is quite particular though, especially for the unexpected decrease between 1995 and 1996[3].

Similarly, the increase in the "other staff" category, which includes all non-medical personnel, might be due to the hiring of administrative staff to deal with the new contracting issues, and/or to the hiring of less qualified personnel to substitute to the nurses.

The software used for the estimations is DEAP 2.1 (Coelli, 1996b). It estimates DEA orientated models and can measure the slack variables using the multi-stage procedure mentioned in Chapter 3. The software also calculates the Malmquist index of TFP, splitting it into its different components. As constant returns to scale will be assumed, for reasons of degrees of freedom, only the general measure of technical efficiency will be taken into consideration, with no distinction between scale and pure technical efficiency.

[3] The suspicion of the value being due to outliers and/or errors in the data was checked for and rejected, as the pattern turned to be common to most units in the sample.

Table 4.1: Summary statistics for outputs: total number of patients treated in every category.

	median	mean (adj. average)	st. dev.	% change*	rate of change**
Inpatients surgery					
1992	3920	5256	4033		
1993	3803	5507	4324	4.8	
1994	4003	5584	4373	1.4	
1995	5021	5585	4424	5.4	
1996	5216	6073	4786	3.2	
1997	5386	5896	4610	-2.9	**2.3**
Inpatients medical					
1992	2516	4004	3543		
1993	2956	4286	3838	7.0	
1994	3575	4172	3643	-2.7	
1995	3205	4374	3963	4.8	
1996	3451	4878	4435	11.5	
1997	3498	5385	4686	10.4	**6.1**
Inpatients others					
1992	2733	3150	3042		
1993	2537	3161	2991	0.3	
1994	2617	3042	2825	-3.8	
1995	2611	2864	2729	-5.9	
1996	1990	2786	2693	-2.7	
1997	1848	2644	2562	-5.1	**-3.4**
Outpatients, day cases, day patients					
1992	73442	102321	88883		
1993	73919	103724	93269	1.4	
1994	76530	119965	108145	15.7	
1995	73822	111872	101561	-6.7	
1996	79158	124616	102702	11.4	
1997	102866	122741	99861	-1.5	**3.7**

** Rate of change between the two adjacent years.*
*** Geometric mean of the yearly changes.*

Table 4.2: Summary statistics for inputs.

	median	mean (adj. average)	st. dev.	% change*	rate of change**
Capital (£000)					
1992	973059	1117942	941880		
1993	913825	1167298	1022602	4.4	
1994	1327970	1296105	1049507	11.0	
1995	1334898	1453971	1177822	12.2	
1996	1180265	1262594	1065260	-13.2	
1997	1380675	1301422	1090141	3.1	**3.1**
Medical staff WTE					
1992	567	606	538		
1993	511	637	585	5.1	
1994	662	651	556	2.2	
1995	747	702	597	7.8	
1996	779	737	632	5.0	
1997	804	812	724	10.2	**6.0**
Nursing staff WTE					
1992	4148	4002	3051		
1993	4113	3828	2875	-4.3	
1994	4260	3896	2781	1.8	
1995	4370	4008	2978	2.9	
1996	3819	3884	2889	-3.1	
1997	3758	3748	2779	-3.5	**-1.3**
Other staff WTE					
1992	2152	2427	1927		
1993	2140	2500	1957	3.0	
1994	3004	2559	1901	2.4	
1995	2986	2764	2136	8.0	
1996	2925	2820	2279	2.0	
1997	3034	2834	2289	0.5	**3.1**
Number of beds					
1992	325	336	247		
1993	317	336	245	0.0	
1994	336	333	230	-0.9	
1995	296	320	229	-3.9	
1996	303	291	206	-9.1	
1997	295	301	222	3.4	**-2.2**

** Rate of change between the two adjacent years.*

*** Overall rate of change per year is calculated as the geometric mean of the percentage changes.*

4.2 Overall change.

The first analysis performed is the calculation of Malmquist indexes of TFP between the two years 1992 and 1997, to have a general picture of the overall change that occurred between the beginning and the end of the reform, that is when more changes were introduced by the new Government.

The results are reported in Appendix 4.1, Table A4.1. These are the unit by unit indexes respectively of the change in technical efficiency (whether the unit has moved closer to the new frontier), of technical progress (which measures the positive or negative shift of the frontier) and finally the change in total factor productivity (the product of the technical change and the total efficiency change).

Pure and scale efficiency are not considered because a constant returns to scale orientation was chosen, for reasons of degrees of freedom. The suspicion was that a variable returns to scale hypothesis would have been too demanding for the given sample size, therefore making the results more questionable.

This was confirmed by the fact that under the VRS hypothesis 80% of the hospitals were deemed as efficient (very little discriminatory power). Consequently, the analysis concentrates on the TFP change as explained by the total efficiency change and the change in technical progress.

Coming to the analysis of the results, Figures 4.1, 4.2 and 4.3 show the distribution of the index values.

Figure 4.1: Frequency distribution of the indexes of efficiency change.

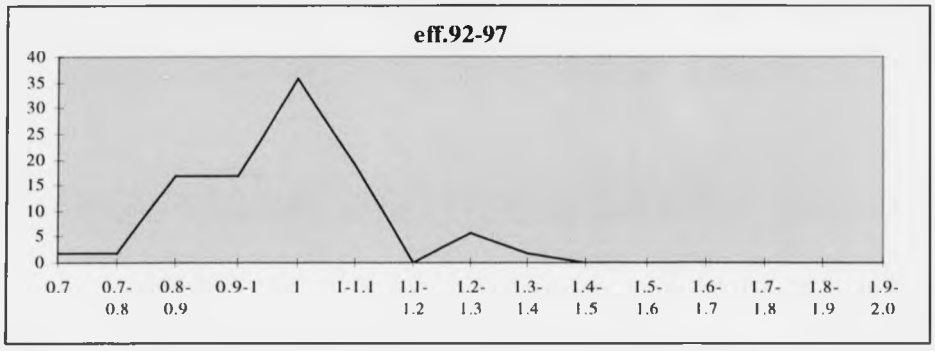


Figure 4.2: Frequency distribution of the indexes of technical progress.

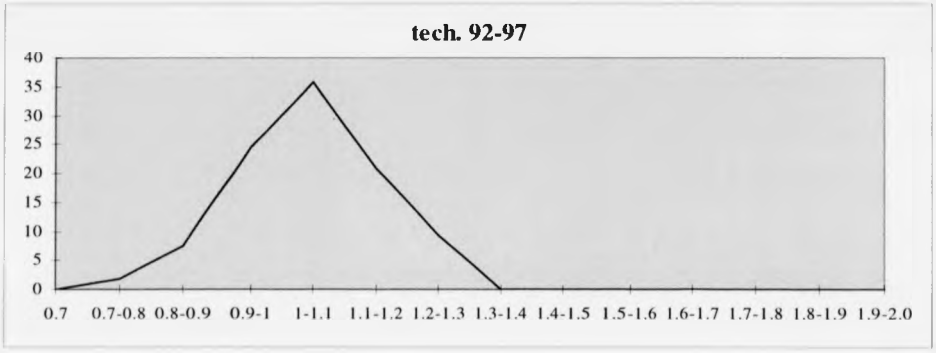
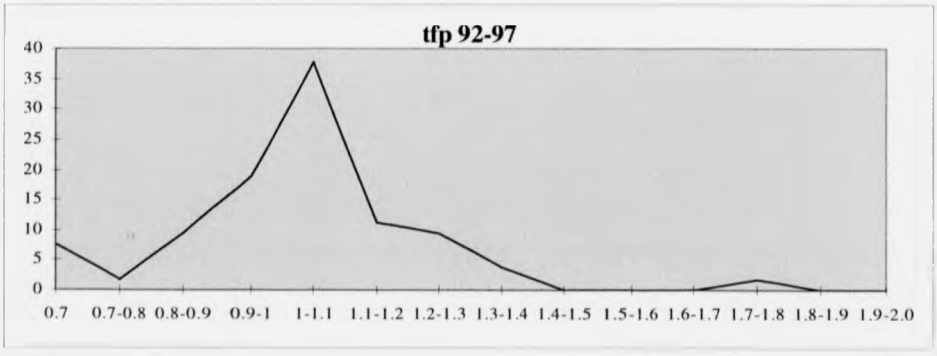


Figure 4.3: Frequency distribution of the Malmquist indexes of TFP.



The values are concentrated around a central point with the exception of very few of them (in general two units, though not always the same ones) that appear as outliers in the distribution. If the average value of the index is calculated, this would be biased by the outliers, and therefore not representative of the average change in hospitals' performance.

In order to have a measure of central tendency, to see whether hospitals have improved or not their performance, an adjusted average is calculated again, by excluding as outliers the observations outside the range $\mu \pm 2\sigma$.

The results are shown in Table 4.3, and more detail is provided in Table 4.4, where hospitals are divided into three categories according to whether they decreased, increased or did not change their performance.

Table 4.3: Malmquist index results, 1992-1997 (adjusted average).

	Adj. average
Efficiency change	0.99
Technical change	1.045
TFP	1.03

Table 4.4: Proportion of hospitals (excluding the outliers) into every category of change.

	Eff.change	**Tech.change**	**TFP**
decreased	37%	34%	37%
same	37%	0%	0%
increased	26%	66%	63%

Table 4.3 indicates an average 3% improvement in hospitals TFP, mainly attributable to a 4.5% improvement in technical possibilities, which is confirmed by the one-to-one correspondence between the improvement in the frontier and the increase in TFP[4].

This means that by adopting new techniques and/or by overall improving the old ones (i.e. keeping the same inputs mix but lowering the ratio of inputs to outputs), hospitals have shifted the output frontier upwards, or the isoquant inwards.

A higher frontier is therefore the probable reason of the worsening of the technical efficiency measure, which scores an average –1%. This results from 37% of the sample being further away from the new, higher, frontier. This negative change is not very big though, and if at the beginning 57% of the sample is efficient, 49% is at the end, with the number of efficient hospitals falling from 30 to 26.

The increase in the number of inefficient units is the main reason of this worsening: the distance from the frontier increases, with the average radial inefficiency going from 0.93 to 0.92, but the average of the *inefficient* hospitals only remains the same (0.84).

Whether the improvement in technical possibilities is due to the adoption of new and better techniques or to a better use of the old ones cannot be said at this stage, as DEA does not estimate a production function. It can however be

[4] See Appendix 4.1, Table A4.1.

deduced from the analysis of the determinants of inefficiency and from the year-by-year changes, as will be seen later on.

4.3 The determinants of inefficiency.

An interesting thing to check is whether the determinants of inefficiency changed in the two years considered. DEA was therefore performed separately for 1992 and 1997, again under the hypothesis of constant returns to scale, and the suggested inputs reductions for the hospitals deemed inefficient were checked.

The calculation was made adding to each unit's radial measure its specific slacks and then averaging these new values out. The results are shown in Table 4.5. Again, an adjusted average measure is preferred because some values can be considered as outliers: not only the deterministic nature of the methodology casts some doubts on calculated wastes of, say, 80%, but also in any case they are much higher than the others.

The columns in the table report the (adjusted) average suggested input reductions for 1992 and 1997 and the difference between the two years; their total average is in the last row.

The average level of inefficiency (i.e. of suggested input reductions) increases between the two periods, though not by a great amount (+1.2%). The ranking instead changes quite a lot, which can be an indication of a change in techniques.

Table 4.5: Average suggested inputs reductions, including the slacks.

	1992	1997	Change
capital	19	25	6
medical staff	15	16	1
nursing staff	18	19	1
other staff	22	27	5
beds	21	14	-7
average	**19**	**20.2**	**1.2**

In 1992 the variable "other staff" and the number of beds are the most wasted inputs among hospitals, followed by capital and nursing staff, whereas the medical staff is (and remains) the least. In 1997 instead capital becomes one of the main determinants of inefficiency, again with the "other staff" variable, whereas the number of beds becomes one of the least wasted ones, and the only input whose use improves instead of worsening over the time period.

One of the effects of the reform, well known also to the general public via the news, was in fact the reduction in the number of beds (see Table 4.2), with consequent ward closures in some cases. This is therefore the probable reason why the input contributes the least to inefficiency at the end, and its calculated waste diminishes.

As regards the capital and "other staff" variables, their levels increase over time, they become the main determinants of inefficiency at the end of the time period and also show the biggest increase in inefficiency level among hospitals.

One interpretation of the above results could be the existence of some substitutability between capital and beds, in the sense that the adoption of more capital intensive forms of treatment reduces the need of overnight stay and therefore of beds. This could be reason of the "switch" in the position of the two inputs as determinants of inefficiency, and could point to a possible "overcapitalisation" by hospitals. Another reason of the increase in the capital inefficiency could be the investment in information technology that hospitals made in order to deal with the new contracting issues (Fattore, 1999), as this would not be directly linked to the treatment of patients. However, the increase in the capital level is also partly due to the change to trust status, which made the hospitals owners of their assets with consequent accountancy changes.

As regards the "other staff" variable, this always contributes most of the inefficiency, but this contribution suffers one of the biggest increases. One reasonable explanation of it could be the increased administrative staff made necessary to deal with the new contracting issues; hospitals were not used to them and it is possible for this change to have been inefficient. Another possibility is that the reduction in nursing staff might have led to the transfer of some of their duties over cheaper but less qualified (therefore more inefficient) staff. A pattern towards the use of cheaper labour inputs in Scottish hospitals was discovered by others (Gray *et al.*, 1986), and this might have been reinforced by the financial concerns of the reform.

The medical staff variable, though recording one of the main increases in level (see Table 4.2) remains as the most efficiently used input across the hospitals, and similarly does the nursing staff.

Some first conclusions can be drawn at this stage. A general comparison between the beginning and the end of the reform time shows hospitals improving their production possibilities and shifting the frontier upwards, by adopting different techniques and/or improving the old ones. The higher frontier, more difficult to reach, increases the number of inefficient hospitals (30 vs. 26) which is the reason of the increase in the overall average inefficiency level, that raises from 7% to 8% respectively without considering the slacks. When only inefficient hospitals are considered, the radial inefficiency measure is 0.84 in both years, that is an average 16% optimal reduction in inputs, and it increases from 19% to 20.2% when also the slacks are considered.

The analysis of the slacks and consequent suggested average input reductions shows an evident change in the contribution of different inputs to inefficiency. This gives more support to the hypothesis that technological change has occurred during time, with hospitals apparently becoming more capital intensive and reducing the use of beds.

To see whether these general considerations are sound it is necessary to look at the year-by-year changes, which is the subject of the next paragraphs.

4.4 Year by year changes.

In this part of the work the efficiency analysis is done for all the years from 1992 to 1997. Some first general information about the performance between 1992 and 1997 can be provided by the separate estimation of DEA for each year in the sample. A summary of the results is provided by Table 4.6[5], which shows for every year the average efficiency score of all hospitals ("efficiency 1", a general performance measure), the average efficiency score of inefficient hospitals only ("efficiency 2"), the number of efficient hospitals, the number of trusts and the number of trusts that are efficient.

Table 4.6: DEA results for each year.

	1992	1993	1994	1995	1996	1997
efficiency 1[1]	0.93	0.97	0.91	0.94	0.93	0.92
efficiency 2[2]	0.84	0.87	0.82	0.84	0.85	0.84
n. eff. units[3]	30	39	27	32	27	26
no. trusts[4]	0	3	19	46	53	53
no. eff. trusts[5]	0	3	10	28	27	26

[1] Average radial DEA measure, calculated over the whole sample.
[2] Average radial DEA measure, calculated for inefficient hospitals only.
[3] Total number of efficient units (i.e. units on the frontier).
[4] Total number of hospital trusts.
[5] Total number of efficient hospital trusts (i.e. hospital trusts on the frontier)

The number of efficient hospitals varies over time, and accordingly so does the general efficiency level ("efficiency 1"), which is the average radial DEA

[5] The complete, unit-by-unit results are in Appendix 4.1, Tables A4.2 to A4.6.

measure calculated over the whole sample. The average "real" inefficiency ("efficiency 2") follows a similar pattern: the values are obviously lower (as only inefficient units are considered), and they reach a minimum in 1994 (0.82, a radial inputs inefficiency of 18%). The last row shows how many of the efficient hospitals are trusts.

When the first wave of change in status takes place in 1994, with 16 new trusts, only 7 are actually 100% efficient. The number of efficient hospital trusts necessarily increases in 1995 as almost the entire sample has undergone the change.

More detailed information about the changes over time is given by the calculation of Malmquist indexes of total factor productivity on all the year pairs. The results are shown in Table 4.7 where again an adjusted average is used as a measure of central tendency for the reasons explained before.

Table 4.7: Malmquist index results (adjusted average), 1992 to 1997.

	Eff.change	**Tech.change**	**TFP**
1992-93	1.026	0.991	1.038
1993-94	0.948	1.075	1.016
1994-95	1.034	0.938	0.966
1995-96	0.985	1.021	1.011
1996-97	0.993	1.017	1.001
geomean	**0.997**	**1.007**	**1.006**

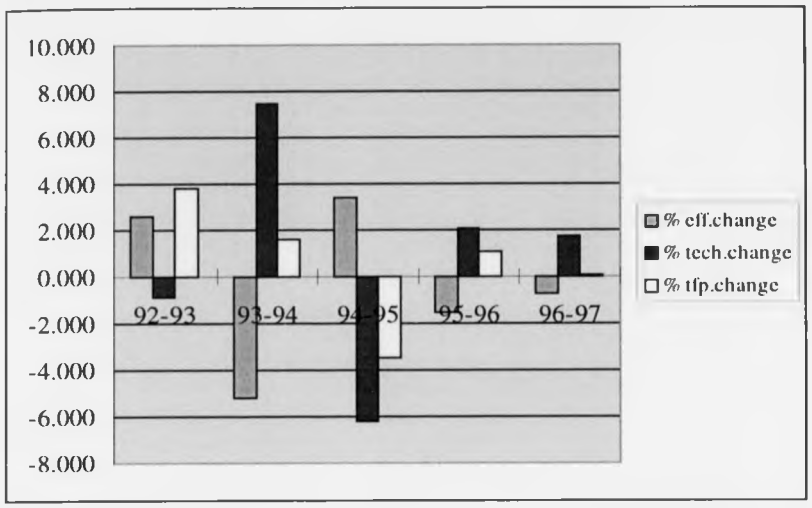
The results indicate a little but quite steady increase in TFP of 0.6% per year, so a total improvement of 3% as before. This results from average shifts of the frontier of a 0.7% a year (3.5% in total) and a worsening of technical efficiency of –0.3% a year (-1.5% in total).

The results are better understood looking at Table 4.8, which reports the number of hospitals in each category of change, and Figure 4.4, that reproduces the same information of Table 4.7 but expressing the changes in percentage terms, to more clearly represent their pattern over time.

Table 4.8: Number of hospitals in each category of change, 1992 to 1997.

		Eff.change	Tech.change	TFP
1992-1993	**decreased**	4	25	16
	same	27	1	1
	increased	22	27	36
1993-1994	**decreased**	25	12	22
	same	25	0	0
	increased	3	41	31
1994-1995	**decreased**	2	43	33
	same	26	0	0
	increased	22	10	20
1995-1996	**decreased**	18	23	27
	same	23	0	0
	increased	12	30	26
1996-1997	**decreased**	17	24	27
	same	23	0	0
	increased	13	29	26

Figure 4.4: Malmquist index results expressed in % terms, 1992 to 1997.



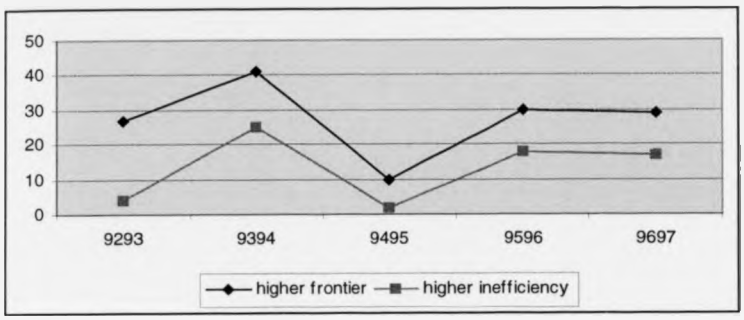
The tables and figure reveal an oscillatory pattern, with major changes taking place at the beginning and then smoothing down towards the end; the shifts of the frontier and the changes in efficiency present such pattern, as they both alternate increases with decreases in every year pair and are also opposite to one another. That is, an improvement in techniques between two years is accompanied by higher inefficiency and especially for 1993/94 and 1994/95 is followed by the opposite phenomenon the next year. This oscillatory pattern translates into a little but steady improvement in TFP, as noted above.

The following interpretation seems appropriate.

As regards the opposite pattern of the shifts of the frontier and the change in technical efficiency, this is attributable to the fact that a higher (lower) frontier is more difficult (easy) to reach. This is confirmed by Figure 4.5, which plots the

number of hospitals that face a positive shift of the frontier against those that worsen their efficiency. The two indexes also have a correlation of -0.85, which further confirms the idea.

Figure 4.5: Number of hospitals facing a higher frontier and number of hospitals worsening their technical efficiency.



The shifts of the frontier itself, and the fact that the oscillatory pattern takes place especially between 1993 and 1994 and between 1994 and 1995, which are the years of the major change to trust status, are interpreted as an indication of a change in techniques, as explained below. This is confirmed also by the inputs' inefficiencies analysis, as will be shown shortly.

Let's recall that the shift of the frontier is an index: it is the average of the indexes measuring the distance between the production possibilities of two years unit by unit, therefore technique by technique, and techniques might change. 1994 is the year in which the first trust wave takes place, with the number of hospital trusts raising from 3 to 19. The hypothesis is that the change to trust status might involve a change in the input mix and/or an adjustment, such that hospitals' productivity could be lower.

When looking at the unit by unit efficiency scores for 1994, it turns out that the majority of hospital trusts is not 100% efficient (see also Table 4.6), which means they are not the main contributors to the positive shift, but they face a higher frontier and score a worsening in efficiency score when compared with the previous year.

In 1995, 27 more hospitals change status, which is half of the sample and raises the number of hospital trusts to 46. If the hypothesis of a change in techniques and a slow down in productivity were correct, given the very high number of new units, this would show in a worsening of the frontier, which is what happens. The lower frontier in turn explains the increase in the efficiency index. Once the major changes have taken place the pattern smoothes down.

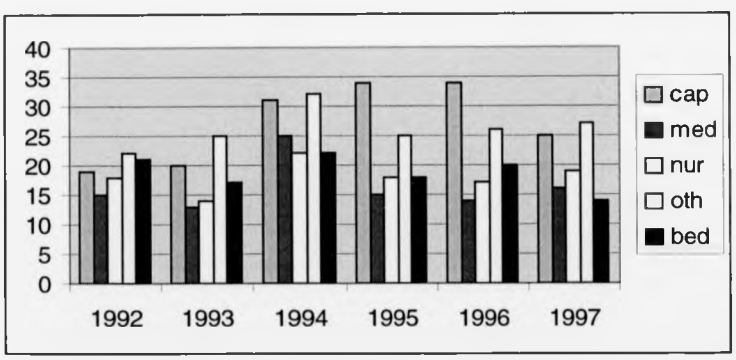
The hypothesis of a change in techniques is confirmed by the analysis of the relative inputs inefficiencies. This is done as previously, by calculating the (adjusted) average inputs reductions taking into account also the slacks, and not only the radial measure. The results are reported in Table 4.9 and shown in Fig. 4.6. The first five rows of Table 4.9 are the suggested inputs reductions when also the slacks are considered, and the last row is their average.

As can be seen in Table 4.9, from 1994 the ranking of inputs changes, with capital and other staff becoming the most wasted inputs and the number of beds becoming instead one of the least wasted ones, similarly to what was observed in the previous section.

Table 4.9: Average % suggested inputs reductions, including the slacks.

	1992	1993	1994	1995	1996	1997
capital	19	20	31	34	34	25
medstaff	15	13	25	15	14	16
nurstaff	18	14	22	18	17	19
otherstaff	22	25	32	25	26	27
beds	21	17	22	18	20	14
average	**19**	**18**	**26**	**22**	**22**	**20**

Figure 4.6: Average % suggested inputs reductions, including the slacks.



It is also interesting to notice (see Figure 4.6) that the average levels are much closer to one another at the beginning and at the end of the time considered, whereas quite a wide variation is shown for the years 1994 to 1996, which are the years of the change to trust status.

This confirms the hypothesis again: the efficient hospitals with respect to which the inefficiencies are calculated are not all trusts (Table 4.6), so they have not all experienced the change in technology, i.e. in the input mix, whereas when all hospitals have changed status the frontier is more homogeneous in terms of techniques, and the average inefficiency levels get closer again (the "efficient peers" all have more similar techniques).

The analysis therefore confirms the idea that technological change has occurred during the time considered, which is associated to the change in trust status. This change in input mix and the probable adjustment to the new situation translates into a slowdown in productivity growth for the units concerned, although overall productivity increases (see section 4.2)[6].

Finally the characteristics of the efficient hospitals are checked for, in particular to see whether the hospitals which first turn into trusts (i.e. in 1993 and 1994) are more efficient than the others. This is a subset of 19 hospitals, as shown in Table 4.6. The non-parametric equivalent to a t-test[7] of the difference of means is conducted (the Mann-Whitney test) to compare their average efficiency over time with the average efficiency of the rest of the sample. The null hypothesis that the means are the same cannot be rejected, not surprisingly as they both have an index of 0.93.

[6] The overall growth in TFP and the shifts of the frontier appear from the overall analysis of paragraph 4.2, as well as from the calculation of Malmquist indexes between different year pairs (1993 and 1995, or 1993 and 1996). Presentation of these last results was considered redundant and therefore not included in the text.

[7] This is necessary because the inefficiency scores are bounded between 0 and 1, so not even asymptotically they could be normally distributed, which is a necessary assumption for the t-test.

The unit-by-unit DEA results show instead a subset of 15 hospitals which are always 100% efficient, only 5 of which change status at the beginning. The test on the equality of means as one might expect this time leads to the rejection of the null hypothesis[8].

It can therefore be concluded that trust status is not significant in determining efficiency and performance improvements, not even when the logic is reversed, as it is not true that most efficient hospitals would turn into trusts first, the others eventually following them.

4.5 Conclusions.

The results of the analysis of this chapter can be summarised as follows.

A general overview of the changes between 1992 and 1997, i.e. the beginning and the end of the reform, shows a 3% improvement in TFP, due to a 4.5% improvement in the frontier. Probably as a consequence of the higher frontier, technical efficiency worsens by –1%, with the number of efficient hospitals reducing from 30 to 26.

The new, higher frontier is characterised by different inputs inefficiencies, a suggestion that the technology of production might have changed, with hospitals moving towards more capital intensive techniques and reducing instead the waste in the use of the beds input.

[8] The tests have been performed both on the average inefficiency over time of each hospital (53 observations) and on the whole vector of inefficiencies (318 observations) and they lead to the

These results are confirmed by the yearly analysis, which concentrates on the changes taking place between each year pair. TFP increases by 3%, again because of improvements in the technology of production (+3.5%), whereas technical efficiency worsens by −1.5%. An opposite oscillatory pattern of the changes in the frontier and those of technical efficiency is revealed, especially in the two years of the major change in status. These years are also characterised by the beginning of the change in technology of production described above, which can therefore be considered a direct effect of the reform: the change to trust status brought with it a change in the way hospitals would provide their services. As will be seen in Chapter 5, also the kind of services provided would change.

However, a link between the new status and a higher efficiency level cannot be proved. Trusts sometimes are on the frontier and sometimes are not, and not even the first wave ones are such because of a better performance compared to the others. The only effect that can be said with certainty is the change in the techniques of production: when introduced the adjustment slows down hospitals' productivity, although overall productivity over time increases.

Naturally, several limitations arise from the very nature of non-parametric estimations, especially from its strong dependency on the actual observations which makes the results less reliable and more sensitive to outliers.

---

very same results.

For this reason, in the next chapter the estimations will be carried using a stochastic, parametric approach. This will enable also to deepen some of the issues related to the characteristics of the technology of production.

APPENDIX 4.1

Table A4.1: Malmquist index results, overall analysis comparing 1992 and 1997.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	0.986	1.1	1.085	28	1	1.009	1.009
2	1	1.143	1.143	29	1	0.937	0.937
3	0.852	1.264	1.078	30	1	1.015	1.015
4	0.898	1.161	1.042	31	1	0.707	0.707
5	0.951	1.168	1.111	32	0.9	0.928	0.836
6	0.842	1.295	1.091	33	1.066	1.133	1.208
7	1.065	1.254	1.335	34	0.847	1.205	1.021
8	1	0.939	0.939	35	0.958	1.082	1.036
9	0.876	0.963	0.844	36	1.038	1.224	1.269
10	0.874	1.097	0.959	37	1.054	0.96	1.012
11	1.06	0.978	1.036	38	1.03	1.037	1.068
12	1	1.02	1.02	39	0.944	1.018	0.96
13	1	1.019	1.019	40	1.202	1.032	1.241
14	0.966	1.07	1.033	41	0.972	1.015	0.986
15	0.887	1.132	1.004	42	1	1.001	1.001
16	1	0.987	0.987	43	1.042	1.044	1.088
17	1	1.139	1.139	44	1.064	1.143	1.216
18	0.802	1.039	0.834	45	1.214	1.086	1.319
19	0.849	0.816	0.692	46	0.641	0.901	0.577
20	1	1.128	1.128	47	1	0.988	0.988
21	1	1.002	1.002	48	1.016	1.02	1.036
22	1	0.942	0.942	49	1	0.991	0.991
23	1.249	1.017	1.27	50	0.98	0.992	0.972
24	0.969	0.846	0.82	51	0.744	0.914	0.681
25	1	1.193	1.193	52	1.961	0.881	1.728
26	1	1.141	1.141	53	1	0.893	0.893
27	1.317	1.671	2.2				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

Table A4.2: Malmquist index results for 1992-1993.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	1.034	1.007	1.042	28	1	1.03	1.03
2	1	1.006	1.006	29	1	1.015	1.015
3	1	1.026	1.026	30	1	1.077	1.077
4	1	1.05	1.05	31	1	0.815	0.815
5	1.002	0.991	0.993	32	1.015	1.016	1.032
6	1	1.048	1.048	33	1.058	0.947	1.002
7	1.065	1.014	1.08	34	1.032	1.054	1.088
8	1	1.044	1.044	35	1.023	1.016	1.04
9	1	0.988	0.988	36	1.095	0.949	1.039
10	1	0.991	0.991	37	1.054	1.013	1.067
11	1.052	1.004	1.056	38	1.03	0.981	1.01
12	1	1.141	1.141	39	1	0.914	0.914
13	1	1.088	1.088	40	0.936	1.028	0.962
14	1.125	0.99	1.114	41	1	1.078	1.078
15	1.073	0.918	0.985	42	1	1.059	1.059
16	1	0.912	0.912	43	1.042	1.035	1.078
17	1	1.058	1.058	44	1.043	1.016	1.06
18	1.399	0.916	1.281	45	1.254	0.923	1.157
19	1	0.829	0.829	46	1.877	0.57	1.07
20	1	1.226	1.226	47	1	0.998	0.998
21	1	0.961	0.961	48	1.3	1.004	1.305
22	0.975	0.973	0.949	49	1	0.813	0.813
23	1.073	0.99	1.062	50	0.924	0.954	0.881
24	1	1.102	1.102	51	0.925	0.885	0.819
25	1	1.054	1.054	52	1.498	0.77	1.153
26	1	1	1	53	1	0.929	0.929
27	1.184	0.913	1.081				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

Table A4.3: Malmquist index results for 1993-1994.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	1	0.996	0.996	28	1	0.852	0.852
2	1	1.022	1.022	29	1	1.04	1.04
3	0.971	0.937	0.909	30	1	0.906	0.906
4	1	1.004	1.004	31	1	1.041	1.041
5	1	1.048	1.048	32	0.825	1.086	0.896
6	1	1.046	1.046	33	0.976	1.121	1.093
7	1	1.047	1.047	34	0.908	1.067	0.968
8	1	0.864	0.864	35	0.9	1.125	1.013
9	0.838	1.158	0.971	36	0.846	1.159	0.98
10	0.851	1.157	0.984	37	0.788	1.342	1.058
11	0.973	1.057	1.029	38	1	1.082	1.082
12	1	0.95	0.95	39	1	1.075	1.075
13	1	1.124	1.124	40	0.93	1.122	1.043
14	0.909	1.126	1.023	41	0.791	0.923	0.73
15	1.046	1.122	1.174	42	1	0.97	0.97
16	1	1.007	1.007	43	1	1.029	1.029
17	1	1.073	1.073	44	0.868	1.043	0.905
18	0.512	1.882	0.964	45	0.747	1.484	1.109
19	0.999	0.924	0.923	46	0.524	1.788	0.937
20	1	1.16	1.16	47	0.924	1.08	0.998
21	1	1.139	1.139	48	1	0.865	0.865
22	0.978	1.026	1.004	49	1	1.211	1.211
23	0.826	1.257	1.039	50	0.954	1.051	1.002
24	1	0.914	0.914	51	0.709	1.347	0.955
25	1	1.132	1.132	52	1.65	1.051	1.734
26	1	1.213	1.213	53	0.977	0.981	0.958
27	0.85	1.269	1.078				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

Table A4.4: Malmquist index results for 1994-1995.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	1	0.996	0.996	28	1	0.891	0.891
2	1	1.017	1.017	29	1	0.943	0.943
3	1.03	1.102	1.135	30	1	0.951	0.951
4	1	1.039	1.039	31	1	0.909	0.909
5	1	1.029	1.029	32	1.017	0.97	0.986
6	1	1.033	1.033	33	1.032	0.964	0.995
7	1	1.043	1.043	34	1.043	0.96	1.001
8	1	0.896	0.896	35	1.051	0.965	1.014
9	0.742	0.915	0.679	36	1.223	0.878	1.074
10	1.175	1.161	1.364	37	1.087	0.845	0.918
11	1.011	0.949	0.959	38	1	0.973	0.973
12	1	0.939	0.939	39	1	1.037	1.037
13	1	0.944	0.944	40	1.183	0.836	0.99
14	1.145	0.924	1.058	41	1.229	0.848	1.041
15	1	0.969	0.969	42	1	1.013	1.013
16	1	1.092	1.092	43	1	0.974	0.974
17	1	0.811	0.811	44	1.176	0.965	1.134
18	1.129	0.646	0.729	45	1.39	0.849	1.181
19	0.94	0.875	0.823	46	1.306	0.711	0.929
20	1	0.792	0.792	47	1.082	0.96	1.039
21	1	0.932	0.932	48	0.914	0.934	0.854
22	1.048	0.864	0.906	49	1	0.723	0.723
23	0.977	0.988	0.965	50	1.031	0.987	1.017
24	1	0.873	0.873	51	1.003	0.789	0.792
25	1	0.964	0.964	52	1.031	0.935	0.964
26	1	0.757	0.757	53	0.986	0.949	0.936
27	1.16	0.878	1.019				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

Table A4.5: Malmquist index results for 1995-1996.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	0.944	1.074	1.014	28	1	1.503	1.503
2	1	1.101	1.101	29	1	0.987	0.987
3	0.84	1.067	0.897	30	1	1.066	1.066
4	0.995	0.995	0.99	31	0.952	0.852	0.811
5	1	0.992	0.992	32	1.058	1.022	1.081
6	0.974	0.995	0.969	33	0.974	1.064	1.037
7	1	1.261	1.261	34	0.924	1.04	0.961
8	1	1.146	1.146	35	0.962	0.989	0.952
9	1.33	0.836	1.112	36	0.856	1.051	0.9
10	0.892	0.899	0.802	37	1.167	1.096	1.28
11	0.988	0.968	0.957	38	1	0.967	0.967
12	1	0.955	0.955	39	0.981	1.007	0.988
13	1	0.98	0.98	40	1.112	0.964	1.072
14	0.848	1.026	0.871	41	1.029	1.051	1.082
15	0.903	1.003	0.905	42	1	0.79	0.79
16	1	0.954	0.954	43	1	1.024	1.024
17	1	1.062	1.062	44	1	1.101	1.101
18	1.27	0.859	1.092	45	1.037	0.919	0.953
19	0.81	0.994	0.806	46	0.602	1.226	0.738
20	1	0.949	0.949	47	1	0.985	0.985
21	1	1.146	1.146	48	0.897	1.317	1.181
22	1	1.218	1.218	49	1	1.362	1.362
23	1.114	0.906	1.01	50	1.101	1.016	1.119
24	0.841	1.109	0.933	51	1.348	1.031	1.389
25	1	0.944	0.944	52	1	1.071	1.071
26	1	1.247	1.247	53	1.038	1.039	1.078
27	1.015	0.981	0.996				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

Table A4.6: Malmquist index results for 1996-1997.

unit	efficiency	tech (shift)	TFP	unit	efficiency	tech (shift)	TFP
1	1.011	0.991	1.002	28	1	0.803	0.803
2	1	1.038	1.038	29	1	1.048	1.048
3	1.014	1.116	1.132	30	1	1.014	1.014
4	0.902	1.06	0.956	31	1.05	1.027	1.079
5	0.949	1.012	0.961	32	0.999	0.99	0.989
6	0.865	1.096	0.948	33	1.027	1.08	1.109
7	1	0.844	0.844	34	0.938	1.061	0.996
8	1	0.993	0.993	35	1.029	1.012	1.041
9	1.059	1.032	1.092	36	1.07	1.17	1.252
10	0.98	1.028	1.007	37	1	0.787	0.787
11	1.037	0.995	1.031	38	1	0.991	0.991
12	1	1.076	1.076	39	0.962	0.95	0.914
13	1	0.979	0.979	40	1.05	0.944	0.991
14	0.972	1.042	1.012	41	0.972	0.982	0.954
15	0.876	1.095	0.959	42	1	1.185	1.185
16	1	0.993	0.993	43	1	0.965	0.965
17	1	1.155	1.155	44	1	1.05	1.05
18	0.781	0.998	0.78	45	0.899	1.026	0.923
19	1.116	1.019	1.137	46	0.828	0.95	0.787
20	1	1.163	1.163	47	1	1.059	1.059
21	1	0.887	0.887	48	0.953	1.007	0.96
22	1	0.96	0.96	49	1	1.082	1.082
23	1.294	0.988	1.279	50	0.98	0.958	0.939
24	1.152	0.951	1.096	51	0.839	0.948	0.796
25	1	1.071	1.071	52	0.77	0.925	0.712
26	1	1.096	1.096	53	1	0.971	0.971
27	1.112	1.628	1.81				

Unit by unit results of the Malmquist index: efficiency change, technical change and total factor productivity change.

CHAPTER 5

THE STOCHASTIC FRONTIER ANALYSIS

The estimations in the previous chapter were carried out using the non-parametric, deterministic DEA. As will be discussed more in detail in Chapter 6, this technique has a few characteristic drawbacks related to its non-stochastic nature:

a) Every distance from the frontier is attributed to inefficiency, as no noise is accounted for.

b) The estimation of the frontier itself is highly sensitive to the presence of outliers, which can therefore bias the results.

Furthermore, by definition the methodology just calculates distances from a frontier with no specification whatsoever of the underlying production function. If this is an advantage as no assumption has to be made a priori about the form of the function itself, on the other hand it limits the analysis of the characteristics and the performance of a sector (inputs and outputs elasticities etc.).

For all these reasons a complementary, stochastic frontier analysis is considered necessary, and this is the subject of this chapter.

As already observed, hospitals are multiple output production units. Whereas this did not create any problems in DEA, the estimation of a regular production frontier as specified and described in Chapter 3 is not possible. Three possible solutions to the problem were discussed then, i.e. the estimation of a cost frontier, the use of index numbers and the estimation of a distance function. The analysis of this chapter consists of the estimation of a stochastic output distance function,

using the model by Coelli and Perelman (1996); index numbers are also used, for reasons that will become apparent in section 5.2.

As regards the choice of a functional form for the deterministic part of the equation, the number of observations makes it possible to estimate a flexible function like the translog and eventually test for more restrictive specifications.

The chapter is organised as follows. Section 5.1 discusses the distance function model. Section 5.2 contains a description of the data and the variables used. The criteria for the model choice are in Section 5.3 and the results and their discussion are in Section 5.4. Section 5.5 covers the issue of technological change and section 5.6 that of the relevance of trust status. The conclusions to the chapter are in Section 5.7.

5.1 The distance function.

In this section, the distance function model proposed by Coelli and Perelman (1996) is presented and then interpreted and discussed.

Let's assume there are N firms that use a vector $x \in R_+^K$ of inputs to produce a vector $y \in R_+^M$ of outputs, and recall from Chapter 3 the definition of an output distance function as a radial expansion measure for the output vector(s) y in order to reach the frontier, i.e.

$$D_o = \min\left\{\vartheta : \frac{y}{\vartheta} \in P(x)\right\} \qquad (5.1)$$

This is homogeneous of degree −1 and weakly monotonically increasing in outputs, and invariant to changes in the units of measurement. The idea in Coelli and Perelman is that (5.1) can be mathematically expressed as a function of the K inputs and M outputs levels of each of the N firms as follows (assuming the log linear, translog function specification):

$$\ln D_{oi} = \alpha_0 + \sum_{m=1}^{M} \alpha_m \ln y_{mi} + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_{mn} \ln y_{mi} \ln y_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} + \\ + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M} \delta_{km} \ln x_{ki} \ln y_{mi} \qquad (5.2)$$

$i = 1,\ldots\ldots, N$
$k = 1,\ldots\ldots, K$
$m = 1,\ldots\ldots, M$

Linear homogeneity in outputs of D_o implies that

$D_o(x, \omega y) = \omega D_o(x, y) \quad \forall \omega > 0$

Hence, one can chose any of the M outputs, say the Mth one, and set $\omega = 1/y_M$ so that

$$D_o(x, y/y_M) = D_o(x, y)/y_M \qquad (5.3)$$

So linear homogeneity can be imposed on (5.2) that becomes

$$\ln\left(\frac{D_{oi}}{y_{Mi}}\right) = \alpha_0 + \sum_{m=1}^{M-1} \alpha_m \ln y^*_{mi} + \frac{1}{2} \sum_{m=1}^{M-1} \sum_{n=1}^{M-1} \alpha_{mn} \ln y^*_{mi} \ln y^*_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} + \\ + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M-1} \delta_{km} \ln x_{ki} \ln y^*_{mi} \qquad (5.4)$$

$i = 1,\ldots\ldots, N$

where $y^*_m = y_m / y_M$. When $y_m = y_M$ the ratio is equal to one, therefore its log is equal to zero and all the terms involving the M-th output disappear from the equation.

For simplicity, let's call TL(.) the whole translog function in (5.4); this can be estimated by noting that

$ln(D_{oi}/y_{Mi}) = TL(.)$

is the same as

$lnD_{oi} - lny_{Mi} = TL(.)$

and therefore

$$-lny_{Mi} = TL(.) - lnD_{oi} \qquad (5.5)$$

Adding a stochastic component $v_i \sim N(0,\sigma^2_v)$ and setting $\ln D_{oi} = -u_i$, equation (5.5) becomes

$lny_{Mi} = -TL(.) + v_i - u_i$

This can be now estimated as a usual production frontier, by regressing (the log of) one output on the (logs of) the inputs and the (logs of) the outputs ratio. Note that the coefficients of a production frontier correspond to the negative of the coefficients of a distance function: so for example a positive elasticity of output with respect to one input corresponds to an elasticity of the distance function with respect to that input which is negative and has the same absolute value. In the one-output case this is pretty intuitive, but the same is true in the multiple output case. Let's specify (5.5) for the multiple output case:

$$\ln y_{Mi} = \alpha_0 + \sum_{m=1}^{M-1} \alpha_m \ln y^*_{mi} + \frac{1}{2} \sum_{m=1}^{M-1} \sum_{n=1}^{M-1} \alpha_{mn} \ln y^*_{mi} \ln y^*_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} +$$
$$+ \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M-1} \delta_{km} \ln x_{ki} \ln y^*_{mi} + v_i - u_i \qquad (5.6)$$

$i = 1,2,\ldots, N$

Equation (5.6) can be viewed as a production function in its own right, in which however the level of one output depends not only on the input levels, but also on the output ratios[1]. The estimated parameters can be interpreted as the parameters of a production function, or as the negative of the parameters of a distance function, exactly as before.

To sum up, in the presence of multiple outputs, a distance function can be estimated. Via the above reparameterisations this is equivalent to estimating a production frontier in which one of the outputs is expressed as a function not only of the inputs, but also of the outputs ratio.

[1] The possibility of simultaneous equation bias resulting from the inclusion of the outputs among the regressors is ruled out by the fact the *ratio* is used, which can be assumed to be exogenous. For details see Coelli and Perelman (1996), page 10.

5.2 The data.

Even though the distance function allows one to overcome the problem of the multiple output, still the number of regressors would have been too high if the translog were estimated using all the 5 output vectors used in Chapter 4.

The number of outputs was therefore reduced from 5 to 2 by constructing two indexes: one for the inpatients and one for the outpatients, day patients and day cases. It was considered more reasonable to keep these two categories separated because they represent very different kinds of treatment, the former having patients spend several days in the hospital and the latter no more than one day, sometimes without even using a bed (and the staff associated to it). As a substitution between the two kinds of services could have taken place, two separate indexes were defined.

The construction of the output index that summarises various categories of treatment makes it necessary to choose some weights that fairly represent the differences among them. The average cost per case and the average length of stay are commonly used.

For these estimations it was decided to use the average cost per case (or category of treatment), on the assumption that more difficult illnesses are more input-demanding than the less serious ones: some might require the use of particular equipment, and/or more medical staff time, as well as a longer time spent in the hospital, which in turn implies more inputs use, and therefore a higher cost.

However, in order not to bias the weights with some measure of inefficiency of each hospital, the average cost per case was calculated as the average cost for the whole of Scotland, and not the average cost per hospital. The index for every hospital was therefore constructed as:

$$y = \sum_{j=1}^{J} q_j c_j \quad j=1,.....,J$$

where q_j is the number of cases treated in each category j, and c_j is the average cost per case in Scotland. As the two main output categories were kept separated, two final output indexes were calculated as above, which are:

y_1 = index of inpatients

y_2 = index of outpatients, day patients and day cases.

The number and kind of inputs are the same as in DEA, so that in the end the variables are

y_1 = index of inpatients

y_2 = index of outpatients, day patients and day cases.

x_1 = capital level (measured in £000)

x_2 = medical staff WTE (whole time equivalent)

x_3 = nursing staff WTE

x_4 = other staff WTE

x_5 = number of beds

The data set is a (balanced) panel of 312 observations, with N= 52 and T=6, that is 52 hospitals[2] observed over a period of 6 years, from 1991-92 to 1996-97[3].

[2] Actually, the number of cross sections was reduced from 53 to 52. As one of the observations had one of the variables equal to 0 at some point in time, and ln0 does not exist, it was preferred to delete that one observation rather than complicating the analysis by using some further transformation, like for example the Box Cox.

[3] From now on, 1991/92 will be referred to as 1992, 1992/93 as 1993 and so on.

5.3 The model.

The model to estimate is the translog output distance function

$$\ln y_{2it} = \alpha_0 + \alpha_1 \ln\left(\frac{y_{1it}}{y_{2it}}\right) + \alpha_{11}\left[\ln\left(\frac{y_{1it}}{y_{2it}}\right)\right]^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} +$$

$$+ \sum_{k=1}^{5}\sum_{l=1}^{5} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln\left(\frac{y_{1it}}{y_{2it}}\right) + v_{it} - u_{it} \qquad (5.7)$$

$i = 1,....,N \; and \; N = 52$

$t = 1,....,T \; and \; T = 6$

over the whole data set[4]. This is equivalent to estimating a single frontier for all the 312 observations, each of which will have a calculated distance from it. A single frontier (with constant parameters across observations and over time) is quite restrictive as an assumption A time effect, represented either with some dummies or a time trend is therefore introduced in the above equation. As discussed in Chapter 3, a time effect might be worth modelling also for the inefficiency component[5]. The specification adopted for the time varying inefficiency is the one proposed by Battese and Coelli (1992), which is automatically performed by the software FRONTIER 4.1:

$$u_{it} = u_i \exp[-\eta(t-T)] \qquad (5.8)$$

As seen in more detail in Chapter 3, a value of $\eta>0$ ($\eta<0$) implies increasing (decreasing) efficiency over time. If $\eta=0$ then there is no time effect, and the hypothesis can be tested by means of a LR test.

---

[4] No dummy for trust status could be introduced for endogeneity reasons. The issue is discussed in detail in section 5.6.

[5] It has to be mentioned at this point that several model specifications were attempted and compared before the final choice was made, including a random effects model and different ML estimations based on different inefficiency distributions but with no time effects. They all presented different problems and were consequently discarded. The software FRONTIER 4.1 was chosen as it allows for a time varying inefficiency, which is not possible with LIMDEP, although the latter can estimate quite a large set of panel data and frontier models. For a more detailed comparison of LIMDEP and FRONTIER 4.1 see Sena (1999).

The procedure followed to select a model was the following.

As before, a model like (5.6) assumes all parameters to be constant over time and across units, so in order to allow for a time effect one can assume either different intercepts for every year or include a time trend. The following models were estimated then: one with 5 dummy variables and an intercept (M1), one with a quadratic time trend (t and t^2) (M2), and one with a linear time trend (M3). These models were tested against each other and against a restricted model with no time effect whatsoever (M4). The test used is the Likelihood Ratio test (LR)[6] which is specified as

$LR = 2[\ \mathcal{L}(H_1)-\mathcal{L}(H_0)] \sim \chi^2_r$

where $\mathcal{L}$ is the value of the maximised log-likelihood and r is the number of restrictions.

In these and all other tests in the chapter, the significance level is always 0.05 unless otherwise stated

The results are shown in Tables 5.1 and 5.2.

Table 5.1: Estimation of models M1 to M4: log likelihood and number of parameters.

	M1	**M2**	**M3**	**M4**
$\mathcal{L}$	178.91	167.81	155.47	151.79
n	33	30	29	28

$\mathcal{L}$ is the value of the maximised log likelihood
n is the number of parameters.

[6] This is because all the tested model pairs are nested in one another. The specification of the restrictions is offered in Appendix 5.1.

Table 5.2: LR tests results (number of restrictions into brackets).

	M1 vs M4	M2 vs M4	M3 vs M4	M2 vs M3	M1 vs M2
LR	54.24 (5)	32.04 (2)	7.36 (1)	24.68 (1)	22.2 (3)
implication	H_0 rejected	H_0 rejected	H_0 rejected	H_0 rejected	H_0 rejected

The 4 models have the following time specifications: M1 = dummy variables (five); M2 = quadratic time trend; M3 = linear time trend; M4 = no time effect.

The tests sequence is the following: first test whether there is a time effect or not: M4 (the null hypothesis) is tested against M1 (dummy variables), M2 (quadratic time trend), and M3 (linear time trend). In all cases the null hypothesis has to be rejected (see Table 5.2), meaning that a time effect exists. Then the three models including a time variable are tested against one another. A quadratic time trend is preferred to a linear one but not to the dummy variables specification, so finally this last one is chosen.

As the parameters' estimates of M1 showed that 2 out of the 5 time dummies are not significant (in particular, those for 1994 and 1995) and have very similar estimates, a different specification in which those two years are put together in a common dummy is attempted next. This therefore translates into a model with 4 time dummy variables:

D_1 = 1 in 1993, 0 else

D_2 = 1 in 1994 and 1995, 0 else

D_3 = 1 in 1996, 0 else

D_4 = 1 in 1997, 0 else.

Call this model M5, it has $\mathcal{L}$ = 178.79 and n = 32; when compared to M1 by means of a LR test the null hypothesis cannot be rejected, so M5 is the final specification used[7].

5.4 The results.

The preceding section showed that the model which best fits the data to measure hospitals' inefficiency is a translog output distance function, in which the logarithm of one of the outputs (in this case y_2, the index of outpatients, day patients and day cases) is regressed on the log of the outputs ratio $y^* = y_1/y_2$ and on the five inputs, with the addition of 4 dummy variables to allow for a different intercept per year.

The equation therefore is

$$\ln y_{2it} = \alpha_0 + \alpha_1 \ln y^* + \alpha_{11}(y^*)^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} + \\ + \sum_{k=1}^{5}\sum_{l=1}^{5} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln y^* + \sum_{t=1}^{4} \zeta_t D_t + v_{it} - u_{it} \qquad (5.9)$$

where

$i = 1,\ldots,N$ *and* $N=52$

$t = 1,\ldots,T$ *and* $T=6$

and

$u_{it} = u_i exp[-\eta(t-T)]$

v_{it} *i.i.d.* $\sim N(0, \sigma^2_v)$

$\varepsilon_{it} = v_{it} - u_{it}$

---

[7] The complete deletion of the two dummies from the equation was discarded on two grounds: the proneness of the translog to multicollinearity and the fact that even if not significant it is

The (in)efficiency component u_{it} is a function of time as shown in (5.9). As regards the distribution of u_i this was first modelled as a truncated normal with mean μ different from 0, i.e.

$u_i = |U_i|$

and

$U_i \sim N(\mu, \sigma_u^2)$ (5.10)

which gave a log likelihood $\mathcal{L}$=178.79. Then it was modelled as a half normal distribution, i.e.

$u_i = |U_i|$

$U_i \sim N(0, \sigma_u^2)$ (5.11)

which gave a log-likelihood $\mathcal{L}$ =178.63. As seen in Chapter 3 the half normal distribution is equivalent to the truncation at 0 of a normal distribution with a 0 mean, so the two models can be compared by testing on (5.10) the null hypothesis H_0: μ=0 against the alternative H_1: μ≠ 0. The LR test of 0.32 led to not reject H_0, and the half normal distribution was chosen, as in (5.11). The results are shown in Table 5.3.

interesting to check the sign of the coefficient when analysing technical change.

Table 5.3: Results of the estimation of equation (5.9), t-ratios into brackets.

parameter	coefficient		parameter	coefficient	
α_0	2.83	(3.36)	β_{24}	-0.12	(-1.30)
α_1	0.11	(0.84)	β_{25}	-0.16	(-0.92)
α_{11}	-0.004	(-0.34)	β_{34}	-0.54	(-2.14)
β_1	0.57	(2.85)	β_{35}	-0.4	(-1.05)
β_2	-0.81	(-2.09)	β_{45}	0.48	(2.42)
β_3	1.9	(2.35)	δ_1	-0.12	(-3.10)
β_4	0.71	(2.04)	δ_2	-0.003	(-0.07)
β_5	-1.66	(-2.48)	δ_3	0.12	(1.38)
β_{11}	-0.04	(-1.89)	δ_4	0.024	(0.46)
β_{22}	-0.10	(-1.83)	δ_5	-0.14	(-1.84)
β_{33}	0.21	(0.7)	ζ_1	0.11	(4.44)
β_{44}	0.14	(1.78)	ζ_2	0.034	(1.16)
β_{55}	0.035	(0.23)	ζ_3	-0.08	(-2.17)
β_{12}	0.10	(1.78)	ζ_4	-0.1	(-2.45)
β_{13}	-0.15	(-1.04)	σ^2	0.115	(4.53)
β_{14}	-0.17	(-1.98)	γ	0.91	(39.2)
β_{15}	0.28	(2.11)	η	0.09	(4.95)
β_{23}	0.5	(2.22)	μ	0	
$\mathcal{L}$	178.63				
OLS $\mathcal{L}$	23.34				

The influence of the inefficiency component is measured by the parameter $\gamma = \sigma^2_u/\sigma^2$, where $\sigma^2=\sigma^2_u+\sigma^2_v$ is the variance of the composite error term $\varepsilon_{it} = v_{it} - u_{it}$ The significance of γ can be tested with an LR test which, if the null hypothesis H_0: $\gamma = 0$ is true, follows a mixed χ^2 distribution[8]. If the null hypothesis is true and inefficiency is not significant, the model is equivalent to a standard "average" production frontier, and its log-likelihood is the same as that of OLS (reported in Table 5.3). The LR test value is 310.58, so the null hypothesis is strongly rejected by the data, meaning that inefficiency is significant.

The significance of η, i.e. of the time effect on (in)efficiency, is similarly tested by means of an LR test. The log-likelihood under the null hypothesis H_0: $\eta = 0$ is 165.04, giving an LR score of 27.2 which again is rejected by the data. Moreover η is positive meaning that inefficiency tends to decrease over time. As

$\partial \ln(u_{it})/ \partial t = -\eta$

the estimated value of 0.09 corresponds to an annual rate of change of u_{it} of 9%. In terms of the distance $D_{oit} = \exp(-u_{it})$, the average scores for each year are reported in Table 5.4. This shows that the average efficiency increases approximately by a 2.5% every year[9].

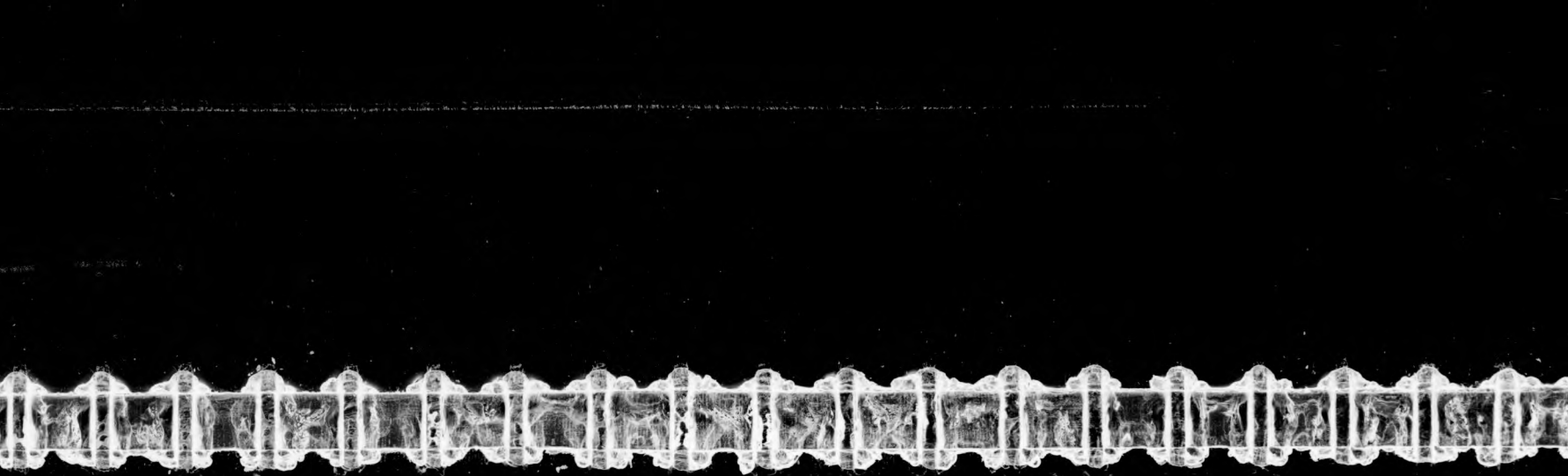
Table 5.4: Average distance values from the estimation of (5.9).

	1992	**1993**	**1994**	**1995**	**1996**	**1997**
D_o (average)	0.695	0.714	0.732	0.750	0.767	0.783

Do is the average value of the output distance function, with $0<D_o\leq 1$: a value of 1 (<1) indicates efficiency (inefficiency).

As regards the significance of the coefficients, the translog function is usually prone to a high level of multicollinearity because of the presence of the squared

[8] See details in Chapter 3.



and interaction terms. This is very often the reason why many of the parameters of translog functions turn out non-significant to the usual t-test even if they are non zero. As a consequence it is preferable not to look at the single t-ratios but to carry out different testing procedures that involve more than one parameter at the same time.

The next step is therefore to calculate the partial and total elasticities and to check for their significance. The partial elasticity with respect to the *k-th* input will be given by

$$e_k = \frac{\partial \ln y_2}{\partial \ln x_k} = \beta_k + 2\beta_{kk} \ln x_k + \sum_{l=1}^{K-1} \beta_{kl} \ln x_l + \delta_k \ln y^* \quad l \neq k \tag{5.12}$$

and the total elasticity is the sum of the partial elasticities.

It is clear from (5.12) that elasticity depends on the inputs and outputs levels, so in order to calculate a general measure for the whole sample all elasticities are calculated at the variables' sample mean.

Six partial elasticities are calculated (five for the inputs and one for the outputs ratio), and a total input elasticity. The results are shown in Table 5.5. The elasticities are calculated over the whole time period as well as for every year separately (in this case using each year's sample means).

Testing for the significance of (5.12) amounts to testing that the whole equation sums to 0. This can be computationally very demanding, but a condition that surely implies this result, and which can be tested, is that all its (seven)

[9] The rate of change of D_0 is $d(\ln D_0)/dt$, and it can be approximated by the difference in the logs when dt=1.

parameters are 0. The results of the LR test are reported in Table 5.6, one for every partial elasticity[10].

Table 5.5: Partial and total elasticities, given the output ratio.

	e_{cap}	e_{med}	e_{nur}	e_{oth}	e_{bed}	e_{tot}	e_y
1992	-0.03	0.47	0.47	0.12	0.11	1.11	-0.74
1993	-0.08	0.46	0.54	0.12	0.05	1.09	-0.75
1994	-0.03	0.48	0.45	0.08	0.17	1.15	-0.75
1995	-0.06	0.46	0.45	0.07	0.21	1.14	-0.74
1996	-0.03	0.44	0.44	0.09	0.23	1.16	-0.73
1997	-0.04	0.45	0.48	0.05	0.21	1.15	-0.72
1992-97	**-0.04**	**0.46**	**0.47**	**0.09**	**0.17**	**1.14**	**-0.74**

Table 5.6: LR test for the significance of the partial elasticities.

	e_{cap}	e_{med}	e_{nur}	e_{oth}	e_{bed}	e_y
LR test	26.68	92.00	26.74	37.26	26.80	410.26
implication	H_0 rejected	H_0 rejected	H_0 rejected	H_0 rejected	H_0 rejected	H_0 rejected

In Table 5.5, the first 5 columns are the partial elasticities of every input, followed by the total input elasticity and then by the elasticity of y_2 with respect to the output ratio $y^* = y_1/y_2$. With the exception of capital, all inputs elasticities are positive, they are all significant and they sum to a total elasticity value of 1.14 (equivalent to mild increasing returns to scale[11] at the sample means), whereas

[10] Another possibility to get round the problem would be to re-estimate the equation normalising the data with their geometric mean. No further detail is provided here because the data set was too big to perform the necessary normalisation procedure.

[11] The constant returns to scale hypothesis is tested by means of an LR test and rejected by the data, as the restricted model has a log likelihood of −13.13.

the elasticity with respect to the outputs ratio, e_y is negative, significant and equal to –0.74. Given the particular functional specification used, e_{tot} measures the effect that an increase in inputs has on the output, given the output ratio: if the output ratio remains the same then a 1% increase in all inputs leads to an increase of 1.14% in both outputs. The effect that an increase in inputs has on the production of y_2 alone, i.e. if the outputs ratio is not kept constant, can be calculated as follows.

Assume for ease of explanation that the estimated function has one input and two outputs (whose ratio is again y^*) and looks like

$$\ln y_2 = \alpha_1 \ln x + \beta_1 \ln y^* + \beta_{11} (\ln y^*)^2 + \gamma \ln x \ln y^* \qquad (5.13)$$

The equivalent to e_y is

$$\frac{\partial \ln y_2}{\partial \ln y^*} = \beta_1 + 2\beta_{11} \ln y^* + \gamma \ln x \qquad (5.14)$$

As y^* is the outputs ratio one can rewrite and then differentiate (5.13) as

$$\ln y_2 = \alpha_1 \ln x + \beta_1 (\ln y_1 - \ln y_2) + \beta_{11} (\ln y_1 - \ln y_2)^2 + \gamma \ln x (\ln y_1 - \ln y_2)$$
$$\partial \ln y_2 = (\alpha_1 + \gamma \ln y^*) \partial \ln x + (\beta_1 + 2\beta_{11} \ln y^* + \gamma \ln x) \partial \ln y_1 - (\beta_1 + 2\beta_{11} \ln y^* + \gamma \ln x) \partial \ln y_2$$

or for ease of interpretation

$$\partial \ln y_2 = A \partial \ln x + B \partial \ln y_1 - B \partial \ln y_2$$

where

$$A = \frac{\partial \ln y_2}{\partial \ln x} \quad \textit{and} \quad B = \frac{\partial \ln y_2}{\partial \ln y^*}$$

so B is the same as (5.14). From (5.14) it is obvious that B has to be ≤ 0. Furthermore, as

$$\frac{\partial \ln y_2}{\partial \ln x} = \frac{A}{(1+B)} > 0 \qquad (5.15)$$

and

$$\frac{\partial \ln y_2}{\partial \ln y_1} = \frac{B}{(1+B)} < 0 \qquad (5.16)$$

it follows that $-1<B\leq 0$; i.e., one has to expect the elasticity of y_2 with respect to the ratio to be negative and in absolute value smaller than 1. The elasticity measure as in (5.16) now represents the relative change in y_2 brought about by a change in y_1, i.e. it is a measure of the substitutability between the two outputs. When B tends to 0 the elasticity will tend to 0, and when B tends to -1 the elasticity will tend to $-\infty$, so that lower absolute values of B imply very little substitutability between the two outputs, and higher absolute values of B a higher substitutability.

The new elasticities as in (5.15) and (5.16) are calculated (again at the variables' sample mean) and the values are reported in Table 5.7.

Table 5.7: Partial and total elasticities, output ratio not fixed.

	e_{cap}	e_{med}	e_{nur}	e_{oth}	e_{bed}	e_y	e_{tot}
Elasticity	-0.15	1.77	1.81	0.35	0.65	-2.85	4.4

Similarly to Table 5.5, Table 5.7 shows that the most productive inputs are the medical and the nursing staff, with elasticities bigger than 1, whereas capital and other staff are least productive with the former showing negative returns. This makes the result difficult to interpret, as it would indicate that capital is always

negatively contributing to production, even before hospitals turned into trusts and eventually started to overcapitalise.

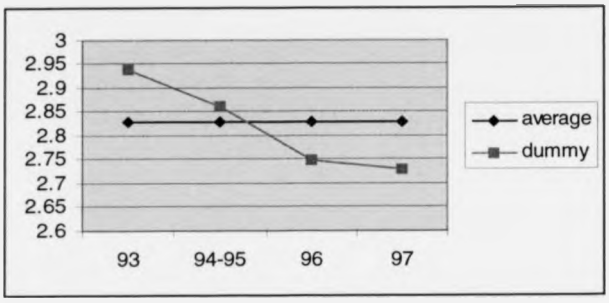
As a consequence it is worth analysing whether this negative value holds true for all the observations in the sample, or if there might have been a change in the technology used by the hospitals during the whole time span, as suggested also in Chapter 4. In other words, because the estimated frontier comes from the pooling of all the observations together, it might be interesting to see whether this is actually correct, or if instead one can detect a change in the production process along time or between particular hospitals categories. This will be done in Section 5.5.

The output substitutability is –2.85, which means that a 1% increase in y_1 (the inpatients) leads to a more than proportional decrease (-2.85%) in y_2 (the outpatients, day patients and day cases). As one might expect inpatients turn out to be more expensive, in terms of resource use, than the outpatients. An inpatient is someone whose treatment requires to spend more than one day in the hospital, which implies the use of more resources as such (bedding and nursing staff) and because of the longer length of stay (one inpatient case lasts for longer than an outpatient case). As the output indexes have been constructed so as to reflect the different resource use in the two cases, the result is even less surprising.

As regards the 4 time dummies they indicate whether a different intercept characterises every year (or year pair), for technological change or other reasons. The intercepts a_t are calculated as $a_t = (\zeta_t + \alpha_0)$, so that a positive dummy parameter means a higher intercept, and viceversa. From Table 5.3, the dummies

are significant with the exception of the one for 1994+1995, and they disclose a pattern around the average (the intercept value) as shown in Figure 5.1.

Figure 5.1: Pattern of the four time dummies estimated in (5.9).



The figure shows that, starting from 1992, there is an increase in 1993 whereas from 1994-95 the trend is decreasing. If the dummies account for technological change then the above results would mean that there is a slowdown in productivity over time, especially after the main change to trust status has taken place. This is not what one would expect, and the very opposite of what seen in Chapter 4, so a different possibility is tested for: the possibility that not only the intercepts, but the very parameters of the equation might have changed over time. This possibility is explored in the next section.

5.5 Testing for technological change.

The estimations carried out in the previous sections pool all the data together and calculate a single frontier which is common to all observations at all points in time. This might be too restrictive a hypothesis, and one might want to test

whether the data can be pooled together in the first place. One possibility of doing this would be to run a series of Chow tests for parameters stability, but this is ruled out for lack of degrees of freedom (the translog has too many parameters to be estimated on a single cross section of 52 observations). An alternative but still valid approach is used instead. This consists of estimating several times the distance function with a time interaction dummy instead of the intercept dummies. In particular, a time dummy d is introduced, which takes a value of 1 for a particular year(s), and 0 else, and this is multiplied to all the variables in the translog distance function, i.e.

$$lny_{2it} = \alpha_0 + \beta' lnx_{it} + d + \rho' lnx_{it} d + \varepsilon_{it} \qquad (5.17)$$

where the x_{it}s are the explanatory variables of (5.9), d is the time interaction dummy and β and ρ are vectors of parameters to estimate. This means that the parameters of the function will be $\beta+\rho$ when $d = 1$, and β when $d = 0$.

The dummy is first set equal to 1 for 1992 (and 0 else), then for 1992 and 1993 (and 0 else) and so on. In this way 5 different distance functions are estimated, each with a different time effect which is captured by the parameters of the interaction dummy. The comparison should shed some light on whether and how things have changed over time. The likelihood results of the five estimations of (5.17) are reported in Table 5.8.

Table 5.8: Log-likelihood of the translog distance function with time interaction dummy.

	92	**92 to 93**	**92 to 94**	**92 to 95**	**92 to 96**
$\mathcal{L}$	177.12	221.84	186.41	184.18	168.16

The significance of the time interaction parameters is tested for by means of LR tests against a restricted model with no time effects and as expected[12] the null hypothesis is always rejected. What is interesting here is to look at how the likelihood of the different specifications changes.

As can be seen in Table 5.8, a much higher value is obtained when separating 1992 and 1993 from all other years. The latter are not very different from one another nor with respect to the value obtained with the varying intercepts model. This points to the fact that the parameters of the distance (and production) function might have changed after 1993.

Before proceeding, the following has to be stressed. A comparison between the above specification and the varying intercepts model can be done using the Akaike information criterion (AIC) for non-nested models. Recall from Chapter 3 that

$AIC = -2\mathcal{L} + 2n$

where n is the number of parameters. When comparing models, the preferred one should be that with the lowest AIC value. Even though on the grounds of the Akaike information criterion the time interaction dummy model should be preferred to the varying intercepts one[13], still it has 56 parameters in it, which is a very high number compared to the 312 observations. This makes the reliability of the choice more fragile; for this reason the results from the pooled, more parsimonious model can still be relied upon for the general analysis, although it is interesting and important to analyse the characteristics of this change in technology.

[12] This is in fact consistent with the results obtained in Section 5.3, when comparing different model specifications.

Given the above, the model in which 1992 and 1993 are separated from the following years is analysed hereinafter. All the results are reported in Appendix 5.2, for reasons of space.

The parameter γ, which measures the relevance of the inefficiency component, is significant (LR test is 343.08). As regards the effect of time on inefficiency, the value of η=0.03 indicates a (significant)[14] improvement, though less marked than in the pooled model. The average increase in efficiency, as measured by the average value of the distance function, in this case is around 0.8% per year, whereas in the pooled model it was 2.5%. Further investigation shows that if the two sub-panels are estimated separately efficiency decreases between 1992 and 1993, whereas it increases between 1994 and 1997. Even though the estimation of 29 parameters on a panel of 52 observations over two years only has not many degrees of freedom, the information is still interesting, raising the question of the relevance of trust status which will be dealt with in the next section.

The estimates from (5.17) are then used to calculate the different elasticities, which are reported in Table 5.9. As seen for equation (5.12), the values are calculated at the variables' sample averages (before and after 1994 in this case) and they are all significant[15].

[13] The time interaction model with a dummy for 1992 and 1993 has an Akaike value of -331.68 and the intercepts model a value of –293.26.
[14] LR test is 5.47.
[15] See Appendix 5.2, Table A5.2.

Table 5.9: Partial and total elasticities, estimation of (5.17).

	e_{cap}	e_{med}	e_{nur}	e_{oth}	e_{bed}	e_y	e_{tot}	e_{y^*}
1992-1993	0.01	0.32	0.37	0.23	0.09	-0.44	1.02	-0.79
1994-1997	-0.02	0.24	0.33	0.08	.089	-0.67	1.52	-2.03

Like in Tables 5.4 and 5.5, e_y is the elasticity of output y_2 (the outpatients, day patients and day cases index) with respect to the outputs ratio y^*, i.e. keeping the latter fixed. e_y^* is the elasticity of output when the output ratio is not constant. e_{tot} is the total inputs elasticity, that is a measure of returns to scale (calculated at the sample mean).

Looking at Table 5.9, all inputs elasticities decrease with the exception of the beds variable. The inputs of capital, other staff and beds show the biggest changes in between the two periods. This is consistent with the results of Chapter 4, and leads to a similar interpretation.

The productivity of capital turns negative, which is a signal of overcapitalisation. This however might also be a consequence of the fact that the acquisition of trust status (which starts in 1994 and covers the whole sample in 1996) involved an increase in the recorded level of capital for accountancy reasons, and it involved investment in information technology for the new contracting issues. As a consequence, concluding for a definite problem of overcapitalisation would be misleading.

The elasticity of "other staff" reduces from 0.23 to 0.08. This reduction can be probably explained by the need to hire administrative staff to deal with the new contracting issues, and/or with the possible use of less qualified personnel for more qualified tasks.

The beds input instead shows a much higher elasticity, which is also the only reason why the elasticity of scale increases. The change in this partial elasticity can be the result of the reduction in the levels of this input[16]. That is, hospitals reduced the number of beds probably far too much even when considering the switch towards treating more outpatients and day patients as opposed to inpatients (the latter require a bed more strictly than the former, but the former still do).

As regards the elasticity of substitution between the two outputs (e_y*) its absolute value increases from 0.79 (close to 1, perfect substitutability) to 2.03. After 1994, if resources are diverted from the treatment of outpatients to that of inpatients the former decrease much more than they used to, with a 1% increase leading to more than a 2% decrease. This is possibly a consequence of the change in treatment patterns that took place over time. Hospitals started reducing the number of beds and increased the proportion of outpatients to inpatients. If in 1992 it was more likely to be required to spend several days in a hospital to undergo some treatment, in more recent years this would happen only for fewer, very serious conditions, which also means more expensive to treat in terms of resource use and in turn explains the massive increase in the elasticity value.

Before concluding this section, an attempt at measuring the change in productivity over time in made. A specific measure as the Malmquist index cannot be use here, but an approximate measure can be given by the changes in the average expected value of the output in each year. The average expected value of lny_2 (the dependent variable) is calculated using the estimates of (5.17) and the variables' sample means for each year. The results are reported in Table 5.10 and show an average increase in productivity of around 9% a year. It has to be kept in mind however that this is an increase in the production of y_2, and the substitution of y_1 with y_2 here is more relevant than ever, as the decrease in the output ratio among the regressors surely affects the result.

Table 5.10: Average expected level of output.

	Average expected output (lny_2)
1992	8.04
1993	8.21
1994	8.38
1995	8.53
1996	8.55
1997	8.48
average rate of change	8.8%

The rate of change is calculated as the difference in the logs.

[16] Cfr Chapter 4, Table 4.2.

What general picture is therefore revealed? After 1994 hospitals start using their resources differently, and their productivity increases, that is the frontier shifts outwards. Both the inputs levels and the way in which these inputs are used (the parameters of the function) do change, and the change is mainly reflected in the fact that they treat patients preferably on a daily basis rather than having them spending more than one day in the hospital: the number of outpatients, day patients and day cases (y_2) in fact increases steadily, whereas some of the categories of the inpatients experience even a negative growth[17].

This strong substitution between the two outputs translates into the high measured annual increase in productivity (shift of the efficient frontier of 8.8%) as this is calculated as the expected value of (the log of) y_2 only. Furthermore, it is also probably the reason why the number of beds is reduced, although the much higher elasticity of the input suggests that the reduction might not be optimal (i.e. the input is still very productive and if increased could contribute to the treatment of many more people). All other inputs have lower elasticity values suggesting that they are used more efficiently.

Finally, efficiency in terms of distance from the frontier also increases: the parameter η (representing the rate of change of the inefficiency component in the error term) is positive and significant, translating into an average efficiency improvement of around 0.8% a year. This value is lower than the 2.5% obtained with the varying-intercepts model, because this did not take technological change properly into account. More specifically, the intuition is this. The varying-intercepts model assumes a common technology, in terms of slope, for all the

[17] Cfr Chapter 4, Table 4.1.

years, "averaging" the characteristics of the technologies used before and after 1994. In this way the units of the first two years are being compared against a frontier that might be emphasising the production of y_2 more than it should, and viceversa for the units after 1994.

This can translate into a larger distance from the frontier in the first two years and a larger improvement in efficiency over time than if the difference in technology is taken into account. The results seem to confirm this intuition: the average distance from the frontier in 1992 is 0.695 in the common technology model, and 0.717 in the other one. In 1997 the order is reversed, with the common technology giving an average distance of 0.783 and the other a value of 0.746.

On the other hand, as already stressed, this limitation of the varying-intercepts model is counterbalanced by the fact that it is much more parsimonious, so that its results can be relied upon for general inference.

5.6 The relevance of trust status

The previous analysis showed a general increase in efficiency and productivity, as well as a change in technology which is very likely related to the change in status. What is interesting to do now is to check whether hospital trusts, so the very working of the reform, are more efficient than non-trusts. That is, whether trust status is a significant variable in explaining efficiency.

No trust dummy was included in the models because of a possible problem of endogeneity. This would be true if hospitals decided to change status because they were more efficient rather than the other way round. However, the possibility was tried, the dummy was non-significant and all the other parameters estimates remained basically the same. Given the endogeneity concern, a different kind of analysis is done.

As was seen in Chapter 3, one could perform a two-stage analysis, where the estimated inefficiencies are regressed on a set of explanatory variables. This however gives less efficient estimates than estimating the parameters of the frontier *and* those of the inefficiency term at the same time. One model of this kind (Battese and Coelli, 1995) can be estimated using FRONTIER 4.1. This consists of modelling the inefficiency component u_{it} as a truncated normal variable coming from a distribution whose mean is in turn a function of a set of explanatory variables, i.e.

$u_{it}=|U_{it}|$

$U_{it} \sim N(m_{it}, \sigma^2)$

and

$m_{it} = \varphi' z_{it}$

where z_{it} is a vector of explanatory variables and φ a vector of parameters to estimate.

For the present case, only one explanatory variable is used in z_{it}, that is a 0-1 dummy for trust status.

This model is first estimated on the whole panel but it fails to converge to a maximum; therefore, two sub-panels, one for the years 1992-1993 and one for 1994-1997, are specified next.

Again, the 1992-1993 panel does not converge to a maximum, probably because of a too low number of hospital trusts then (3 out of 104 observations). The estimation for 1994-1997 does instead work and gives a value of $\mathcal{L}$=76.29; the trust dummy's parameter is estimated as negative but non-significant (t ratio = -1.44).

To double-check on the result a second stage regression can be tried, in which the estimated inefficiencies from (5.17) are regressed on a trust dummy. Even if less efficient than the simultaneous estimation carried out above, it has the advantage of including all the years.

One concern could be that by the way inefficiency was modelled, the ranking of the units is the same every year because it is based on the value of u_{it} at time T. However this calculation is made using the information of the whole panel data set, that is considering the behaviour of units through time[18]. It is therefore still informative to carry out the estimations.

The model consists of regressing the whole vector of 312 inefficiencies on separate trust dummies, as shown in (5.18) below. In this way one can test two things: whether trust status is in general a relevant factor in explaining (in)efficiency (through the joined significance of the parameters); and whether

[18] In other words this means that the inefficiency scores and the ranking of the panel data set are not the same of those of a frontier that uses only data at time T.

the trusts of a particular year are significantly different from other hospitals (through the single parameter's significance).

The model is specified as

$$\hat{u}_i = \alpha_0 + b_1 t_1 + b_2 t_2 + b_3 t_3 + \varepsilon_i \qquad (5.18)$$

where

t_1= 1 if trust in 1993, 0 else

t_2 = 1 if trust in 1994, 0 else

t_3 = 1 if trust in 1995, 0 else

Only years 1993, 1994 and 1995 are in the equation, because in 1992 no hospitals are trusts, therefore making a vector of 0s, and in 1996 and 1997 all of them are, thus giving vectors of 1s, and the matrix would be singular.

The results of the estimation of equation (5.18) are reported in Table 5.11. The very low R^2 (0.003) is explained by the fact that none of the dummies' parameters is significant, neither jointly nor separately.

Table 5.11: Results of the estimation of (5.18).

parameter	estimate	t-ratio	F test
α_0	2.102	95.87	
b_1	0.02	0.11	
b_2	0.09	1.03	0.356
b_3	0.004	0.07	
R^2	0.003		

The results are therefore confirmed: trust status is not significant in explaining efficiency scores, neither is there any evidence that the first trusts were more efficient than other hospitals.

5.7 Conclusions.

In this chapter, a stochastic output distance function has been estimated in different ways, to measure and analyse the changes in efficiency and productivity of 52 acute hospitals in Scotland in the period between 1991/92 and 1996/97.

First, a general pooled model shows all hospitals improving their efficiency over time at a rate of 2.5% a year. One frontier only is calculated for all the hospitals, with fixed parameters though varying intercepts to allow for technological change. The unexpected negative pattern of change revealed by the intercept dummies raised the possibility that the slope parameters of the equation, and not just the intercepts might have changed over time.

A different model specification is therefore estimated, with a time interaction-dummy which allows for the change in parameters. In this way a structural break is detected that separates the technologies used before and after 1994, which is also when hospitals start changing status. The new technology shows very different inputs elasticities from the older one and is more directed towards the treatment of patients on a day basis, which translates into a higher opportunity cost of treating an inpatient. Efficiency increases over time, now by a 0.8% a year, and productivity, measured as the average expected value of the dependent

variable, increases quite significantly (8.8% a year). This last measure could be used as an approximation of the shift of the frontier over time, but by construction it will emphasise only one the two output vectors.

The switch to a different technology and the change of status almost coincide, as the first big trust wave takes place in 1994, but trust status is not revealed as relevant in determining efficiency.

It is plausible to conclude that over time performance has improved, both in terms of shifts of the frontier itself and of distances from it, and hospitals have changed the technology used for the provision of their services as well as the kind of services provided, but that this improvement is specific to hospital trusts is not proved. Without data covering the period before the introduction of the reform, it cannot be excluded that such improvements could be a general pattern of change over time.

How these results compare with those obtained in Chapter 4, and what the general picture shows, are the subject of the next chapter.

APPENDIX 5.1

Selection of models M1 to M4 estimated in Section 5.3.

The comparison of models in Section 5.3 was made by means of LR tests because in each case one model was nested in the other. The restrictions imposed in each case are specified hereinafter.

For all the models:
N=52
T=6
NxT=312
f(.) is the translog distance function, with n = 27 parameters

Model 1: 5 intercept dummy variables

$$M1 = \alpha_0 + f(.) + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4 + \delta_5 D_5 + \varepsilon_{it}$$

where

D_1=1 for time 2 and 0 else
D_2=1 for time 3 and 0 else
D_3=1 for time 4 and 0 else
D_4=1 for time 5 and 0 else
D_5=1 for time 6 and 0 else

Model 2: quadratic time tend

$$M2 = \beta_0 + f(.) + \beta_1 t + \beta_2 t^2 + \varepsilon_{it}$$

Model 3: linear time trend

$$M3 = \gamma_0 + f(.) + \beta_1 t + \varepsilon_{it}$$

Model 4: no time effect

$$M4 = \varphi_0 + f(.) + \varepsilon_{it}$$

The restriction(s) imposed in the tests were the following:

1) M1 vs M4
   M4 is nested in M1 if
   $\delta_1=\delta_2=\delta_3=\delta_4=\delta_5=0$
   number of restrictions:5

2) M2 vs M4
   M4 is nested in M2 if

$\beta_1=\beta_2=0$
number of restrictions:2

3) M3 vs M4
M4 is nested in M3 if
$\beta_1=0$
number of restrictions:1

4) M2 vs M3
M3 is nested in M2 if
$\beta_2=0$
number of restrictions:1

5) M1 vs M2
M2 is nested in M1 if
$\delta_3 = \delta_2 - \delta_1$
$\delta_4 = 2\delta_2 - 3\delta_1$
$\delta_5 = 3\delta_2 - 4\delta_1$
number of restrictions: 3

This comes from observing that M1 could be reparameterised as
$\alpha_0 = \beta_0 + \beta_1 + \beta_2$
$\delta_1 = \beta_1 + 3\beta_2$
$\delta_2 = 2\beta_1 + 8\beta_2$
$\delta_3 = 3\beta_1 + 15\beta_2$
$\delta_4 = 4\beta_1 + 24\beta_2$
$\delta_5 = 5\beta_1 + 35\beta_2$

APPENDIX 5.2

Results from the estimation of equation (5.17)

Table A5.1: Parameters' estimates of equation (5.17). Standard errors into brackets.

parameter	coefficient		parameter	coefficient	
α_0	3.81	(4.25)	β_{13}	0.09	(0.53)
α_1	0.64	(3.39)	β_{14}	-0.23	(-2.51)
α_{11}	-0.06	(-2.61)	β_{15}	0.22	(1.37)
β_1	0.37	(1.66)	β_{23}	0.36	(1.45)
β_2	-0.21	(-0.47)	β_{24}	-0.01	(-0.07)
β_3	1.01	(1.14)	β_{25}	-0.24	(-1.22)
β_4	0.83	(2.26)	β_{34}	-0.51	(-1.80)
β_5	-1.55	(-2.25)	β_{35}	-0.09	(-0.25)
β_{11}	-0.07	(-2.48)	β_{45}	0.38	(1.66)
β_{22}	-0.07	(-1.14)	δ_1	-0.10	(-2.17)
β_{33}	0.04	(0.12)	δ_2	0.01	(0.08)
β_{44}	0.14	(1.91)	δ_3	-0.03	(-0.22)
β_{55}	0.002	(0.01)	δ_4	0.14	(2.03)
β_{12}	0.05	(0.86)	δ_5	-0.18	(-1.73)

This part of the table shows the value of the parameters when the dummy is equal to 0.

parameter	coefficient		parameter	coefficient	
ρ_1	0.03	(0.21)	ρ_{15}	0.16	(1.19)
ρ_2	-5.58	(-3.85)	ρ_{16}	-0.44	(-1.63)
ρ_3	-0.57	(-2.17)	ρ_{17}	0.07	(0.43)
ρ_4	-0.01	(-0.17)	ρ_{18}	0.04	(0.19)
ρ_5	0.7	(1.21)	ρ_{19}	0.30	(0.99)
ρ_6	-3.02	(-4.47)	ρ_{20}	0.31	(1.44)
ρ_7	-0.97	(-0.81)	ρ_{21}	0.11	(0.47)
ρ_8	2.99	(3.43)	ρ_{22}	-1.00	(-1.91)
ρ_9	1.50	(1.61)	ρ_{23}	-2.75	(-0.40)
ρ_{10}	0.03	(0.66)	ρ_{24}	0.36	(0.88)
ρ_{11}	-0.26	(-2.96)	ρ_{25}	0.11	(1.27)
ρ_{12}	0.80	(1.47)	ρ_{26}	0.001	(0.02)
ρ_{13}	-0.04	(-0.24)	ρ_{27}	0.20	(1.22)
ρ_{14}	-0.24	(-0.89)	ρ_{28}	-0.22	(-1.90)
$\mathcal{L}$	**221.84**		σ^2	**0.18**	(4.8)
OLS $\mathcal{L}$	**50.30**		γ	**0.96**	(87.74)
			η	**0.03**	(2.21)

This part of the table shows the parameters of the interaction dummy. The order of the parameters is the same as in the first part of the table.

Table A5.2: LR tests results on the significance of the elasticities from the estimation of equation (5.17). Number of restrictions into brackets.

	e_{cap} **LR test**	e_{med} **LR test**	e_{nur} **LR test**	e_{oth} **LR test**	e_{bed} **LR test**	e_{y} **LR test**
1992-1993 (7)	20.12	47.1	31.6	25.88	48.84	466.06
1994-1997 (14)	26.84	106.4	47.66	77.44	62.88	498.88

The null hypothesis is rejected in all cases, so all elasticities are significant at the 5% level.

Table A5.3: Average distance values from the estimation of (5.17).

	1992	**1993**	**1994**	**1995**	**1996**	**1997**
D_o (average)	0.717	0.723	0.729	0.735	0.741	0.746

Do is the average value of the output distance function, with $0<D_o\leq 1$: a value of 1 (<1) indicates efficiency (inefficiency).

CHAPTER 6

COMPARISON AND CONCLUSIONS

The aim of this chapter is to finally put together all the results of the thesis and draw some general conclusions. The comparison of the techniques used for the empirical analysis and their results is the subject of Section 6.1, and the general conclusions of the whole work are in Section 6.2.

6.1 Conclusions from the empirical analysis.

In order to compare the results obtained in Chapters 4 and 5 it is useful to briefly explain what are the differences in the two techniques used (DEA and the econometric stochastic frontier) and the reasons why they were used both[1].

DEA is a non-parametric, deterministic approach that by linear programming constructs a piecewise linear frontier using the information contained in the data. The actual levels of inputs and outputs are observed unit by unit; to measure efficiency, every unit is "compared" with, or contracted against, the closest supporting hyperplane, i.e. with the units that have the most similar technique, in terms of inputs to outputs ratios.

The non-parametric nature means that no assumption is made about the existence of a production function common to the observations, which eliminates the risk of

[1] See for example Cubbin and Tzanidakis (1998), and Drake and Weyman Jones (1996).

misspecification of the production technology. Units can be very different in the way they produce their output, and in this case each unit will always be compared with the ones most similar to it. This comparison translates into specific targets of inputs reduction or outputs increase for every (inefficient) observation. This detail on the single units of the sample makes the methodology better suited when this is the focus of the analysis. An additional advantage of the non-parametric nature is that DEA can readily handle multiple inputs and multiple outputs, and is fairly easy to estimate.

DEA is deterministic because no allowance is made for any statistical noise. The lack of distributional assumptions, which is a sense could be an advantage, is the major limitation of the method. First, the frontier is constructed on the basis of the actual observations, which makes it extremely sensitive to outliers. Furthermore, the lack of statistical noise means that every distance from this frontier is attributed to inefficiency, whereas it could be due to other factors (measurement errors, factors beyond the control of the firm etc.). This in turn can lead to an inaccurate measurement of inefficiency, both overestimating the efficiency of some exceptional units and/or underestimating the efficiency of others.

The second problem is that the lack of distributional assumptions eliminates the possibility of proper statistical testing of hypotheses. Non parametric tests can be used, which are however weaker than the parametric ones. One of the consequences of it is that the choice of the variables becomes crucial: the decision has to be made beforehand not only of what factors are expected to be relevant but also in which way they are expected

to be so. Different variables can lead to very different results, but a choice cannot be made on proper statistical grounds, making the approach less reliable.

The econometric stochastic frontier is almost the very opposite to the above, as it is parametric and stochastic. Inefficiency is modelled as an asymmetrically distributed component of the error term. The stochastic nature makes the results more reliable, in themselves and because the models can undergo proper statistical tests.

As a parametric approach, assumptions have to be made about the behaviour and objectives of the units in the sample. However, the specific problem of having to choose a functional form can be reduced by the possibility of testing different models, and with large enough data sets flexible functional forms can be used to minimise the number of restrictions.

Other assumptions are necessary, on the distribution of both the stochastic and the inefficiency component of the error term, and this can affect the results as discussed in Chapter 3. Finally, multiple inputs and multiple outputs are not as easy to manage as they are in the non-parametric approach.

It appears from the above that the characteristics, and so the strengths and weaknesses of the two methods, are quite complementary. By not imposing any functional restriction DEA can give quite detailed information about the units in the sample; when used to

calculate a Malmquist index it translates into specific measures of TFP change and its components. Its main shortcoming lies in its deterministic nature, as discussed above.
SF provides instead a general picture of the characteristics of a sector, with the possibility of measuring elasticities and other production characteristics. It is better at giving a general, statistically more reliable description of the technology and the general direction of changes over time. However, depending on the model estimated, it can be less good at separating the TFP components.

As a consequence, the use of both techniques would strengthen the analysis of a sector, as they reinforce each other through their comparison and the emphasis on different aspects.

Coming to the specific models estimated in Chapters 4 and 5 a few differences have to be stressed, which are the reason why, apart from what said above about their opposite strengths and focus, a direct comparison in terms for example of ranking of units is not appropriate. These differences can be summarised as follows (the reasons have been discussed in Chapters 4 and 5):

1) The unit of measurement of output is different: 5 output vectors are used in DEA, 2 indexes are used in SF.
2) DEA imposed constant returns to scale, whereas SF revealed increasing returns to scale.
3) The number of cross sectional units was reduced from 53 to 52 for SF.

This makes it inappropriate to make a one-to-one comparison. It is more meaningful to see what general story the two analyses tell, if the results are similar and if not why this is the case and if there is any room for reconciliation.

First of all, both techniques point towards a general improvement of productivity over time, and towards a change in the technology of production. Hospitals appear to have changed the way in which they provide their services, with a possible excess of capital and other staff and a reduced use of beds.

SF shows that this change is not only significant, but is associated to a change in the kind of services provided, with a marked switch to the treatment of patients on a day-basis. This change appears to be associated to the change in status. The structural break in SF is 1994, when the first trust wave takes place, and the result is confirmed by the DEA - Malmquist results: inputs inefficiencies start changing then, and this translates into the oscillatory pattern of adjustment between technical efficiency and technological change observed more in detail in Chapter 4. If the change in status brought with it a change in what hospitals produce and how they produce it, however the very fact of being a trust does not appear to be synonymous of higher efficiency. Both chapters concluded for the non relevance of being a trust to explain efficiency scores.

The main difference between the results lies in the quantification of the productivity improvement and in its attribution to either technological progress or technical efficiency.

DEA calculates an improvement in TFP of 3% overall, which is attributed to shifts of the frontier, whereas technical efficiency seems to worsen by a –0.3% a year, so –1.5% overall (using the year by year results for ease of comparison with SF). SF instead concludes for an improvement in technical efficiency over time of a +0.8% a year when allowing for the change in technology, and for very pronounced shifts of the frontier.

There are two main reasons for this difference.

First of all, output has been measured in different ways. If Malmquist indexes are calculated again using the two output vectors of SF the change in efficiency is now a +0.4% a year[2]. Furthermore, SF showed the existence of increasing returns to scale whereas DEA assumed constant returns to scale; as seen in Chapter 3 this can lead to an overestimation of the inefficiency of units. The risk that inefficiency might have been overestimated by DEA is even greater if the deterministic nature of the approach leads to an overestimation of the efficient frontier.

So, if on the one hand the improvement could be a consequence of the output unit of measurement used in SF (unnecessarily complicated for DEA which allows for multiple outputs) on the other hand the negative change could be a consequence of the deterministic nature of this approach. A straightforward conclusion in this respect is therefore not possible, except in the sense that if technical efficiency has changed over time, this change has not been very pronounced.

[2] These results are not included in the thesis as this last analysis was done only to confirm the interpretation.

As regards the measurement of technological progress, the model specification of Chapter 5 does make it rather unsuitable to the aim. The model points to the existence of a time effect, and the fact that this turns out to be a change in the slope parameters casts a lot of doubts on the effect captured by the intercept dummies.

Furthermore, Chapter 5 estimated a *distance* function, with the outputs ratio among the regressors. This is interesting because it allowed to reveal and measure the substitution effect between the two main categories of output, but limits the possibility of correctly measuring the shifts of the frontier over time. Only the rate of increase in y_2 could be calculated, besides by a regression that has a decreasing ratio y_1/y_2 on the RHS.
Given this limitation of the SF model, the DEA and Malmquist index results are better suited to draw conclusions about the shifts of the frontier.

Putting all the above together, it finally seems appropriate to conclude in this way.

The productivity of hospitals has improved over time in terms of shifts of the efficient frontier and possibly, though much more moderately, in terms of distance from it. No direct link could however be proved to exist between the status of trust and a higher efficiency, as was expected instead by the reform. Another direct effect is instead proved, which is the change in what hospitals produce and how they produce it. More people are treated on a day-basis, capital levels increase, the number of beds is reduced and a shift towards less costly staff probably takes place.

What answer can be given to the two questions posed at the beginning of the thesis is therefore discussed in the next section.

6.2 Conclusions of the thesis.

This thesis performed an analysis of the effects of the 1989 NHS reform of hospital services. Without privatising them, this reform introduced competition by creating separate figures of purchasers and providers which would operate on the "internal market" on the basis of contractual relationships. The idea behind the reform was that competition would have improved the efficiency of the provision, leading to reduced waste of resources, higher quality levels and more choice to the patients. No real guidance was given by the law about the kind of contracts to use, but the general aim was to move towards forms of prospective payment based on the average cost of the treatments.

The theory of contracts analysed in Chapter 2 showed that defining an "optimal" contract for hospital services, if at all possible, is at the very least extremely difficult. Various models have been developed by the literature, and one was developed in Section 2.3 to deal in particular with the issue of waiting time. Beyond the specific differences among these models, one thing that seems common to most of them is the recognition not only of the complexity of the issue, but also of the potential drawbacks of prospective payment systems. Apart from the informational requirements, the problems in the identification and measurement of output and the potentially significant transaction costs, the literature

showed that this kind of contracts give incentives to cost minimisation, and therefore technical efficiency, but can lead to sub-optimal decisions especially (but not exclusively) with respect to quality.

The results of the empirical analysis seem not only to confirm such concerns, but to raise others. The improvement in technical efficiency is not proved for sure, and in any case is not very marked. Moreover, no link is revealed between trust status and efficiency.

The main improvement in productivity over time seems to be given by the shifts of the frontier itself, i.e. by technological change. This change in technology is shown to be not only a change in the way hospitals provide their services, but also in the kind of services they provide.

This, in our opinion, could be a reason of concern.

A trend is revealed in which patients are treated more and more on a day basis. This could be the result of technological progress, i.e. the investment in specialised equipment and staff, but in that case the elasticity of the input would not be negative and/or its "waste" recorded by DEA would not increase. Such results suggest that a lot of the increase in capital levels could be due to investments related to the new contracting issues. This is confirmed also by the changes in the measured waste and the elasticity of the "other staff" variable. In such a scenario, the reduction in the length of stay and the treatment of patients on a day basis could actually imply a reduction in the quality level

of the service provided. As already stressed, no direct measure of quality was available for this thesis, but the concern is not ill founded and the conclusion supported by others' results (Maniadakis *et al.*, 1999).

The new financial concerns therefore seem to have led hospitals on the one hand to divert resources towards the new contracting activity, and on the other hand to reduce the amount and possibly the quality of the resources devoted to treatment: by using less qualified staff, and by reducing the length of treatment. This therefore puts the measured increase in productivity in a different light.

There are obviously limitations to the present work, which open the way to future research. As seen in Chapter 2, some of the assumptions of the theoretical model could be relaxed in order to make the framework more complex and realistic. Empirically, it would be interesting to identify and use some measure of quality. Furthermore, the availability of data covering the period before the introduction of the reform would be particularly interesting: this would in fact allow to make more precise and direct comparisons in the trends of productivity and efficiency change, giving more sound grounds to the above conclusions.

Bibliography

Afriat, S. (1972), "Efficiency estimation of production functions", *International Economic Review*, 13(3), pp. 568-598.

Aigner, D., Chu, S. (1968), "On estimating the industry production function", *American Economic Review*, 58, pp. 826-839.

Aigner, D., Lovell, C. A. K., Schmidt, P. (1977), "Formulation and estimation of stochastic frontier production function models", *Journal of Econometrics*, 6, pp. 21-37.

Ali, A. I., Seiford, L. M. (1993), "The mathematical programming approach to efficiency measurement", in Fried *et al*, op. cit.

Allen, R., Gertler, P. (1991), "Regulation and the provision of quality to heterogeneous consumers: the case of prospective prices of medical services", *Journal of Regulatory Economics*, 3, pp. 361-375.

Appleby, J. (1994), *Developing contracting: a national survey of DHAs, boards and NHS trusts*, NAHAT, Birmingham.

Baggott, R. (1994) *Health and Health Care in Britain*, St.Martin's Press.

Banker, R. D., Charnes, A., Cooper, W. W. (1984), "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science* 30(9), pp. 1078-1092.

Barker, K., Chalkley, M., Malcomson, M.J., Montgomery, J. (1996), "Contracting in the NHS; legal and economic issues", *Discussion Papers in Economics and Econometrics*, no. 9622, University of Southampton.

Bartlett, W., Le Grand, J. (1992), "The impact of NHS reforms on hospital costs", *Studies in decentralisation and quasi-markets,* n.8, SAUS, University of Bristol.

Bartlett, W., Harrison, L. (1993), "Quasi markets and the NHS Reforms", in Barlett *et al.*, op.cit.

Barlett, W., Le Grand (1994a) *Quasi-markets and Social Policy*, MacMillan.

Bartlett, W., Le Grand, J., (1994b), "The performance of trusts", in Robinson *et al.*, op. cit.

Battese, G. E., Corra, G. (1977), "Estimation of a production frontier model: with application to the pastoral zone of eastern Australia", *Australian Journal of Agricultural Economics*, 21, pp. 167-179.

Battese, G. E., Coelli, T. J. (1992), "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India", *Journal of Productivity Analysis*, 3, pp.153-169.

Battese, G. E., Coelli, T. J. (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data", *Empirical Economics*, 20, pp. 325-332.

Besley, T., Hall, J., Preston, I. (1999), "The demand for private health insurance: do waiting lists matter?"" *Journal of Public Economics*, 72, pp. 155-181.

Burns, P., Huggins, M., Riechmann, C., Weyman-Jones, T. (2000), "Benchmarking the Dutch electricity network utilities", DTe commissioned report.

Butler, J. (1994) "Origins and early development", in Robinson, R. and Le Grand, J., op.cit.

Butler, J. (1995) *Hospital Cost Analysis*, Kluwer Academic Publishers.

Cave, D.W., Christensen, L.R., Tretheway, M.W. (1978), "Flexible cost functions for multiproduct firms", *Review of Economics and Statistics*, pp.477-481.

Caves, D. W., Christensen, L. R., Diewert, W. E. (1982a), "Multilateral comparisons of output, input and productivity using superlative index numbers", *Economic Journal*, 92, pp. 73-86.

Caves, D. W., Christensen, L. R., Diewert, W. E. (1982b), "The economic theory of index numbers and the measurement of input, output and productivity", *Econometrica*, 50, pp. 1393-1414.

Chalkley, M., Malcomson, M.J. (1995a), "Contracting for health services with unmonitored quality", *Discussion Papers in Economics and Econometrics*, no. 9510, University of Southampton.

Chalkley, M., Malcomson, M.J. (1995b), "Contracting for health services when patient demand does not reflect quality", *Discussion Papers in Economics and Econometrics*, no. 9514, University of Southampton.

Chalkley, M., Malcomson, M.J. (1995c), "Contracts and competition", *Discussion Papers in Economics and Econometrics*, no. 9513, University of Southampton.

Chalkley, M., Malcomson, M.J. (1996), "Contracts for the National Health Service", *Discussion Papers in Economics and Econometrics*, no. 9641, University of Southampton.

Chalkley, M., Malcomson, M.J. (1998), "Government purchasing of health services", *Discussion Papers in Economics and Econometrics*, no. 9821, University of Southampton.

Charnes, A., Cooper, W. W., Rhodes, E. (1978) "Measuring the efficiency of decision making units", *European Journal of Operational Research*, 2(6), pp.429-444.

Charnes, A., Cooper, W. W., Seiford, L. M., Stutz, J (1982), "A multiplicative model for efficiency analysis", *Socio-Economic Planning Sciences*, 16(5), pp. 223-224.

Charnes, A., Cooper, W. W., Seiford, L. M., Stutz, J. (1983), "Invariant multiplicative efficiency and piece-wise Cobb-Douglas envelopments", *Operations Research Letters* 2(3), pp. 101-103.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L. M., Stutz, J. (1985), "Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions", *Journal of Econometrics*, 30, pp. 91-107.

Charnes, A., Cooper, W. W., Rousseau, J., Semple, J. (1988), "Data envelopment analysis and axiomatic notions of efficiency and reference sets", Research report CCS 558, Centre for Cybernetic Studies, University of Texas.

Charnes, A., Cooper, W. W., Lewin, A., Seiford, L. M. (1994), *Data Envelopment Analysis: Theory, Methodology and Application*; Kluwer Academic Publishers.

Cobb, S., Douglas, P. (1928), "A theory of production", *American Economic Review*, 18, pp. 139-165.

Coelli, T. J. (1995), "Estimators and hypothesis tests for a stochastic frontier function: a Monte Carlo analysis", *Journal of Productivity Analysis*, 6, pp. 247-268.

Coelli, T. J. (1996a), "A guide to FRONTIER version 4.1: a computer program for stochastic frontier production and cost function estimation", *CEPA working paper* 96/07, University of New England, Australia.

Coelli, T. J. (1996b), "A guide to DEAP version 2.1: A data envelopment analysis computer program", *CEPA working paper* 96/08, University of New England, Australia.

Coelli, T. J., Perelman, S. (1996), "Efficiency measurement, multiple output technologies and distance functions: with application to European railways", *CREPP working paper* 96/05, Université de Liège.

Coelli, T. J. (1998), "A multi-stage methodology for the solution of orientated DEA models", *Operations Research Letters*, 23, pp. 143-149.

Coelli, T. J., Rao, D. S. P., Battese, G. E. (1998), *An Introduction to Efficiency and Productivity Analysis*; Kluwer Academic Publishers.

Cooper, W. W., Seiford, L. M., Tone, K. (2000), *Data Envelopment Analysis. A Comprehensive Text with Models, Applications, References and DEA-Solver Software*; Kluwer Academic Publishers.

Conrad, F.R., Strauss, R.P. (1983), "A multiple-output, multiple-input model of the hospital industry of North Carolina", *Applied Economics*, 15, pp. 341-352.

Cowing, T.G., Holtmann, A.G. (1983), "Multiproduct short-run hospital cost functions: empirical evidence and policy implications from cross-section data", *Southern Economic Journal*, pp.637-653.

Cubbin, J., Tzanidakis, G. (1998), "Regression versus data envelopment analysis for efficiency measurement: an application to the England and Wales regulated water industry", *Utilities Policy*, 7, pp. 75-85.

Culyer, A.J. (1971), "The nature of the commodity health care and its efficient allocation", *Oxford Economic Papers*.

Culyer, A.J. (1991), *The Economics of Health*, E.Elgar Publ.

Debreu, G. (1951), "The coefficient of resource utilisation", *Econometrica* 19(3), pp.273-292.

Deprins, D., Simar, L., Tulkens, H. (1984), "Measuring labour efficiency in post offices", in Marchand *et al.*, op.cit.

Deprins, D., Simar, L. (1989), "Estimation de frontières déterministes avec facteurs exogènes d'inefficacité", *Annales d'Economie et de Statistiques*, 14, pp. 177-150.

De Fraja, G. (2000), "Contracts for health care and asymmetric information", *Journal of Health Economics*, 19, pp. 663-677.

DoH, (1989), *Working for Patients*, HMSO, London (Cm 555).

Drake, L., Weyman-Jones, T. (1996), "Productive and allocative inefficiencies in U.K. building societies: a comparison of non-parametric and stochastic frontier techniques", *Manchester School of Economic and Social Studies*, 64, pp. 22-37.

Ellis, R.P., McGuire, T. (1986), "Provider behaviour under prospective reimbursement", *Journal of Health Economics*, 5, pp. 129-151.

Ellis, R.P., McGuire, T. (1991), "Optimal Payment systems for health services", *Journal of Health Economics*, 9, pp.375-396.

Ellis, R.P. (1998), "Creaming, skimping and dumping: provider competition on the intensive and extensive margins", *Journal of Health Economics*, 17, pp. 537-555.

Elwood, S. (1996), "Full-cost pricing rules within the NHS", *Management Accounting Research*, 7, pp. 25-51.

Evans, R.G. (1971), "Behavioral cost functions for hospitals", *Canadian Journal of Economics*, 4, pp.198-215;

Evans, R.G. (1981), "Incomplete vertical integration: the distinctive structure of health care industry", in van der Gaag, J., Perelman, M., (1981), *Health, Economics and Health Economics*, Amsterdam, North Holland;

Färe, R., Lovell, C. A. K. (1978), Measuring the technical efficiency of production", *Journal of Economic Theory*, 19(1), pp. 150-162.

Farrell, M. J. (1957), "The measurement of productive efficiency", *Journal of the Royal Statistical Society*, Series A, General, 120(3), pp. 253-281.

Fattore, G. (1999), "Cost containment and health care reforms in the British NHS", in Mossialos et.al., op.cit.

Feldstein, M.S. (1967), *Economic Analysis for Health Service Efficiency*, North Holland Publishing Company.

Fisher, I. (1922), *The Making of Index Numbers*, Boston, Houghton Mifflin.

Fournier, G.M., Mitchell, J.M. (1989), "Hospital costs and competition for services: a multiproduct analysis", *The Review of Economics and Statistics*, pp. 627-634.

Fried, H. O., Lovell, C. A. K., Schmidt, S. S. (1993), *The Measurement of Productive Efficiency- Techniques and Applications*, Oxford University Press.

Gabrielsen, A. (1975), "On estimating efficient production functions", *Working Paper A-85, Chr. Michelsen Institute*, Department of Humanities and Social Sciences, Bergen, Norway.

Ganley, J.A., Cubbin, J. (1992), *Public Sector Efficiency Measurement*, Amsterdam; London: North-Holland.

Gray, A., McGuire, A., Stuart, P. (1986), "Factor input in NHS hospitals", *Discussion Paper* n.2, University of Aberdeen.

Greene, W.H. (1980), "Maximum likelihood estimations of econometric frontier functions", *Journal of Econometrics*, 13, pp. 27-56.

Greene, W.H. (1990), "A gamma distributed stochastic frontier model", *Journal of Econometrics*, 46, pp. 141-163.

Greene, W.H. (1997), "Frontier production functions", in Pesaran *et al.*, op. cit.

Greene, W. H. (2000), "Simulated likelihood estimation of the normal-gamma stochastic frontier function", mimeo, Stern School of Business, New York University.

Grosskopf, S. (1993), "Efficiency and productivity", in Fried *et al.*, op.cit.

Hoffmeyer, U.K., McCarthy, T.R. (1994), *Financing Health Care*, vol.1, Dordrecht, London, Kluwer.

Holliday, I., 1992, *The NHS reformed*, Baseline Books.

Holmstrom, B., Milgrom, P. (1991), "Multitask principal-agent analyses: incentive contracts, asset ownership and job design", *Journal of Law, Economics and Organisation*, 7, pp.24-52.

ISD Scotland, Scottish Health Service Costs, NHS in Scotland, 1991/92-1996/97.

Koopmans, T. C. (1951), "An analysis of production as an efficient combination of activities" in T. C. Koopmans, ed., *Activity Analysis of Production and Allocation*, Cowles Commission for Research in Economics, no.13; J. Wiley & Sons, Inc., New York.

Kumbhakar, S. C., Ghosh, S., McGuckin, J. T. (1991), "A generalised production frontier approach for estimating determinants of inefficiency in US dairy farms", *Journal of Business and Economic Statistics*, 9(3), pp. 279-286.

Kumbhakar, S. C., Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*, Cambridge University Press.

Laffont, J.J. Tirole, J. (1993), *A Theory of Incentives in Procurement and Regulation,* MIT Press, Cambridge, MA.

Le Grand, J. (1994), "Evaluating the NHS Reforms", in Robinson, *et al.*, op.cit.

Lovell, C. A. K. (1993), "Production frontiers and productive efficiency", in Fried *et al.* (1993), op. cit.

Ma, C.A. (1994), "Health care payment systems: cost and quality incentives", *Journal of Economics and Management Strategy*, 3, pp. 93-112.

Malmquist, S. (1953), "Index numbers and indifference surfaces", *Trabajos de Estatistica*, 4, pp. 209-242.

Maniadakis, N., Thanassoulis, E. (1997), "Changes in the productivity of a sample of Scottish hospitals: a cost index approach", *Research Paper* no. 288, Warwick Business School.

Maniadakis, N., Hollingsworth, B., Thanassoulis, E. (1999), "The impact of the internal market on hospital efficiency, productivity and service quality", *Health Care Management Science*, 2, pp. 75-85.

Marchand, M., Pestieau, P., Tulkens, H. (1984), *The Performance of Public Enterprises: Concepts and Measurement*. Amsterdam, North Holland.

Meeusen W., Van den Broek, (1977), "Efficiency estimations from Cobb-Douglas production functions with composed error", *International Economic Review*, 18, pp. 435-444.

McGuire, A., Westoby, R. (1983), "A production function analysis of acute hospitals", *Discussion Paper* n. 4, University of Aberdeen.

McGuire, A. (1985), "Methodological considerations of hospital production and cost functions: relationships to efficiency", *Discussion Paper* n.8; University of Aberdeen.

McGuire, A., Henderson, J., Mooney, G. (1988), *The Economics of Health Care*, Routledge and Kegan, London;

Mossialos, E., Le Grand, J. (1999), *Health Care and Cost Containment in the European Union*, Ashgate, USA.

Newhouse, J.P.(1970), "Towards a theory of non-profit institutions: an economic model of a hospital", *American Economic Review*, 60, pp. 64-74.

Olesen, O. B., Petersen, N. C. (1995), "Chance constrained efficiency evaluation", *Management Science* 41, pp. 442-457.

Parkin, D., Hollingsworth, B. (1997), "Measuring production efficiency of acute hospitals in Scotland 1991-1994: validity issues in data envelopment analysis", *Applied Economics*, 29, pp. 1425-1433.

Pauly, M., Redisch, M. (1973), "The not-for-profit hospital as a physicians' cooperative", *American Economic Review*, 63, pp. 87-99.

Pesaran, M. H., Schmidt, P. (1997), *Handbook of Applied Econometrics*, vol.2, Microeconomics, Blackwell.

Propper, C. (1994), "Quasi-markets, contracts and quality in health and social care: the US experience", in Bartlett *et al.*(1994a), op.cit.

Propper, C. (1995), "The disutility of time spent on the UK's National Health Service waiting lists", *Journal of Human Resources*, 30, pp. 677-700.

Propper, C. (1996), "Market structure and prices: the responses of hospitals in the UK NHS to competition", *Journal of Public Economics*, 61, pp. 307-335.

Propper, C.(2000), "The demand for private health care in the UK", *Journal of Health Economics*, 19, pp. 855-876.

Richmond, J. (1974), "Estimating the efficiency of production", *International Economic Review*, 15, pp. 515-521.

Robison, R. and Le Grand, J. (1994), *Evaluating the NHS reforms*, The King's Fund Institute.

Rogerson, W. (1994), "Choice of treatment intensities by a non profit hospital under prospective pricing", Journal of Economics and Management Strategy, 3, pp. 7-51.

Shephard, R. W. (1953), *Cost and Production Functions*, Princeton N.J.; Princeton University Press.

Shephard, R. W. (1970), *Theory of Cost and Production Functions*, Princeton N.J.; Princeton University Press.

Schmidt, P. (1976), "On the statistical estimation of parametric frontier production functions", *Review of Economics and Statistics*, 58(2), pp. 238-239.

Schmidt, P., Sickles, S. (1984), "Production frontiers and panel data", *Journal of Business and Economic Statistics*, 2, pp. 367-374.

Scott, A., Parkin, D. (1995), "Investigating hospital efficiency in the new NHS: the role of the translog cost function", *Health Economics*, 4, pp. 467- 478.

Sena, V. (1999), "Stochastic frontier estimation: a review of the software options", *Journal of Applied Econometrics*, 14, pp.579-586.

Smith, K., Wright, K. (1994), "Principals and agents in social care", *York Working Papers*, n.123, University of York.

Söderlund, N., Csaba, I., Gray, A., Milne, R., Raftery, J. (1997), "Impact of the NHS reforms on English hospital productivity: an analysis of the first three years", *British Medical Journal*, 315, pp.1126-1129.

Stevenson, R. (1980), "Likelihood functions for generalised stochastic frontier estimation", *Journal of Econometrics*, 13, pp. 58-66.

Tatchell, M. (1983), "Measuring hospital output: a review of the case mix and service mix approaches", *Social Science and Medicine*, vol. 17, pp.871-883.

Törnqvist, L. (1936), "The Bank of Finland's consumption price index", *Bank of Finland Monthly Bulletin*, 10, pp.1-8.

Tulkens, H., Vanden Eeckaut, P. (1990), "Productive efficiency measurement in retail banking in Belgium", *CORE Working Paper*, Université Catholique de Louvain, Belgium.

Wagstaff, A.(1988), "Econometric studies in health economics: a survey of the British literature", *Journal of Health Economics*, pp.1-51.

Wagstaff, A. (1989), "Estimating efficiency in the hospital sector: a comparison of three statistical cost frontier models", *Applied Economics*, 21, pp.659-672.

Weyman-Jones, T. (2001), "Stochastic non-parametric efficiency measurement in electricity distribution", Loughborough University (mimeo).

DX

223806