Efficiency and productivity analysis in regulation and governance

ISSN 1750-4171

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

Efficiency and Productivity Analysis in Regulation and Governance

Thomas Weyman-Jones

WP 2007 - 02

Dept Economics
Loughborough University
Loughborough
LE11 3TU United Kingdom
Tel: + 44 (0) 1509 222701
Fax: + 44 (0) 1509 223910
http://www.lboro.ac.uk/departments/ec

Efficiency and Productivity Analysis in Regulation and Governance

Thomas Weyman-Jones

Department of Economics, Loughborough University, UK

t.g.weyman-jones@lboro.ac.uk

Abstract This paper surveys the application of efficiency and productivity analysis to recent regulatory experience, especially in Europe. From a review of regulatory case studies, particularly of network industries, it is clear that regulatory practice differs from theoretical precedent in choice of methodology, sample size, model specification and price or revenue control implementation. A principal-agent model of linear regulatory contracts is used to understand this discrepancy, suggesting that efficiency and productivity analysis has been used to capture economic rent rather than to provide incentives for efficiency. Predictions of the model are used to investigate other assumptions in efficiency and productivity analysis.

Keywords: regulation, data envelopment analysis, stochastic frontier analysis

JEL classification: D24, L25, L51

North American Productivity Workshop, *NAPW4*
Stern School of Business, New York University,
June 2006

Efficiency and Productivity Analysis in Regulation and Governance

Thomas Weyman-Jones*

June 2006

Abstract

This paper surveys the application of efficiency and productivity analysis to recent regulatory experience, especially in Europe. From a review of regulatory case studies, particularly of network industries, it is clear that regulatory practice differs from theoretical precedent in choice of methodology, sample size, model specification and price or revenue control implementation. A principal-agent model of linear regulatory contracts is used to understand this discrepancy, suggesting that efficiency and productivity analysis has been used to capture economic rent rather than to provide incentives for efficiency. Predictions of the model are used to investigate other assumptions in efficiency and productivity analysis.

Keywords: regulation, data envelopment analysis, stochastic frontier analysis

JEL classification: D24, L25, L51

1. Introduction

This paper[1] asks the question: in the application of efficiency and productivity analysis to regulated industries and government agencies, why do regulators not behave more like the text-books say they should? More prosaically, it has two purposes: to illustrate the issues that regulators have become aware of in applying efficiency and productivity analysis, and to determine if experience of regulatory practice has lessons for theoretical and empirical developments in efficiency and productivity analysis.

It is useful to survey these issues using the experience of some of the data envelopment analysis and econometric studies carried out by European regulators and Governments in the last two decades[2] . A fundamental question is why governments and regulators are interested in efficiency and productivity analysis. Three different objectives are possible: the aim may be simply to be able to measure efficiency or productivity in regulated entities or government agencies in order to be better informed; alternatively, the aim may be to reinforce the impact of incentives for greater efficiency, or, finally, the aim may be to address the issue of capturing economic rent. I shall argue that in the area of governance, i.e. management of government agencies, the first objective is most important. In the area of regulation of investor-owned firms, it is natural to suppose that, aside from simple measurement, the principal objective of regulatory use of efficiency and productivity analysis is to reinforce incentives. However, I shall argue that this is mistaken, and that the principal role is to determine the efficient transfer of economic rent from regulated producers to consumers, and society as a whole. This will be apparent from the nature of regulatory games. Surprisingly, this objective gives a

*Department of Economics, Loughborough University, UK, email address: t.g.weyman-jones@lboro.ac.uk

much more important role to efficiency and productivity analysis than the focus on high-powered incentives.

2. Evaluating recent regulation experience

In recent years, comparative efficiency and productivity analysis, more commonly ‘benchmarking’, has become widely used by European, Latin American and Australian network regulators as part of these regulatory regimes. This has gone hand in hand with the wide adoption of forward looking incentive regulation often based on price-capping or revenue-capping. The European Union Electricity Directive of 2003 requires that ex ante regulation will become the norm throughout the European Union, leading to wider use of efficiency and productivity analysis as noted by Filippini, Farsi, and Fetz (2005). Such regulation involves the setting of a regulatory contract for subsequent years usually with a price or revenue cap and a supporting comparative efficiency analysis.

The general form of such price or revenue caps relates the upper limit on the regulated price or total revenue to a base price or total cost base adjusted each year by the rate of retail price inflation (RPI) minus an X-factor or productivity offset.

$$Price\ Cap : P = P_0\,(1 + RPI - X) \tag{1}$$

$$Revenue\ Cap : R = C_0\,(1 + RPI - X) \tag{2}$$

Scandinavian countries and the UK were among the first to apply efficiency and productivity analysis techniques to regulating network industries, CEPA (2003a). Sweden has used both non-parametric procedures and model network analysis (MNA) to determine yardsticks of performance, while Finland and Norway have used DEA explicitly in regulatory design for both rate of return regulation and revenue capping, as described in Viljainen et al (2004). Subsequently, the Netherlands explicitly adopted DEA in setting price caps, and several Australian State regulators such as the NSW Independent Price and Regulatory Tribunal, have followed suit with increasingly more complex modelling procedures, IPART (1999), Carrington et al (2002). Regulators usually state that efficiency and productivity analysis is integral to yardstick competition, and that the motivation for benchmarking is the concern for incentives: e.g. Jamasb and Pollitt (2003:1609) interpret regulators’ stated objectives as: “The aim is to provide the regulated firms with an incentive to close their efficiency gap with the frontier firms”. This is partly done by publishing the results of the analysis, and partly by incorporating the results explicitly into the determination of the regulations for the price or cost base (P_0, C_0) and the X-factor in the price cap or a revenue cap. In pursuing this objective, regulatory authorities have endeavoured to adopt procedures that are "ideally supported by academic experts in the field", Hargreaves et al (2006: 24), and Bundesnetzagentur (2006).

The different regulatory case study evaluations have been able to identify four key issues that recur frequently. These are

1. the choice of methodology
2. the nature of the sample of firms
3. the specification of the models, particularly the variables to be used
4. the translation of the results into the regulatory mechanism

Methodology

CEPA (2003a, 2003b), for example, notes that DEA, the parametric programming approach, DFA, SFA, total factor productivity by growth accounting, and DEA based Malmquist productivity growth measures have all been considered by regulatory agencies. In practice, however the range of methodologies used has been much narrower with DEA and COLS from cross-section samples playing major roles in many of the recent regulatory studies. Viljainen et al (2004) review seven case studies in electricity distribution and note that three used DEA, three used model network analysis, and one used COLS. Filippini, Farsi, and Fetz (2005) also comment on the prevalence of DEA and COLS in electricity. In the UK, COLS has been the preferred methodology for the two major networks with a history of price control reviews, water and electricity. Wherever SFA has been tried the results have been unusable due to the limited sample sizes employed, CEPA (2003a), Pollitt (2005).

Sample size

Regulatory reviews have demonstrated little willingness to look beyond cross-section samples of the national regulated firms in the year of the review. This is despite the availability of large panel data sets or international comparators, Carrington et al (2002). For example, both the UK water and electricity regulators have access to firm level data for periods of 10 to 15 years but in each of their price reviews have only examined cross-section samples of the small number of regulated firms under their jurisdiction: around 25 in the case of UK water supply and 14 in the case of UK electricity. In Scandinavian countries the sample sizes are potentially much larger with a prevalence of local municipal networks, but again cross-section sampling is preferred and the additional problem that these producers (unlike the investor-owned-utilities) do not have the option of rejecting the regulatory contract is ignored.

Specification

In the majority of cases studied, operating expenditure (OPEX) has been the dominant concern of the regulatory efficiency and productivity analysis. This has been the case in all UK electricity reviews, in all but one of the UK water reviews and in all but one of the Scandinavian cases cited by Filippini, Farsi, and Fetz (2005). The key issue seems to the difficulty of defining a relevant measure of total cost (TOTEX) for the utilities in question. Partly this is a reporting issue arising from laxity in defining regulatory accounting procedures, and partly it arises because of the confusion between an economic model of the regulated firm and the corporate finance model that is often preferred by regulators. In an economic model of the firm, total cost is simply the value which the firm's inputs would have in comparable activities undertaken by other firms, i.e. total expenditure on volume measures of inputs (including capital services) at competitive market input prices. In the typical regulatory model, total expenditure is defined as:

$$TOTEX = OPEX + DEPRECIATION + RETURN\ ON\ CAPITAL \quad (3)$$

Here RETURN ON CAPITAL is given by the real weighted average cost of capital (WACC) multiplied by the value of assets relevant to the regulated activities or regulatory asset value (RAV). RAV itself is dynamically changing due to the effect of capital expenditure, CAPEX:

$$RAV_t = RAV_{t-1} + CAPEX - DEPRECIATION \quad (4)$$

In comparing TOTEX across different regulatory jurisdictions these definitions can lead to non-comparable data, not least because of differences in measuring RAV. Viljainen et al (2004) report a variety of book value calculations used in Scandinavian countries, while UK regulators have used equity market valuations of RAV soon after privatisation to set the base figure. To overcome this, in their comparison of international frontiers, Jamasb and Pollitt (2003) had to use OPEX + CAPEX as a current cash flow measure of TOTEX. Coelli et al (2003) and Diewert and Lawrence (2002) offer clear guidance on capital measurement issues but these have not been widely adopted by regulators.

Translation to price control

The benchmarked costs from an efficiency and productivity analysis of regulated firms is then used to generate both the price or total cost base and the X-factor, but again there seems no settled procedure for doing this, Burns Jenkins and Riechmann (2005). Some regulators generate firm-specific X-factors as in the case of UK water or the proposals for German electricity distribution, while some generate a single X-factor for the industry with firm specific base prices (UK electricity).

The evaluation of nearly two decades of efficiency and productivity analysis in regulation especially in Europe suggests that if the purpose is to define widely approved and transparent incentive procedures, then many regulators are failing to set basic standards of comparison. There is a lack of clearly defined cost measures, a reluctance to adopt efficient statistical procedures or econometric models which are grounded in efficiency and productivity analysis , a preference for small cross-section samples when international panel data are available, and a variety of different ways of signalling the incentive levels which vary among regulatory jurisdictions. This should make us examine the role of efficiency and productivity analysis more closely.

3. Efficiency and productivity analysis in incentive regulation

It can shown that yardstick competition does not necessarily require an analysis of the efficient frontier – if the incentive mechanism is designed to ensure that the regulated firm will minimise cost in its own sub-game, the level of the incentive is immaterial.

Shleifer (1985) uses the average of the observed company costs but excluding the cost of the firm in question to determine the regulated price: $p_i = \overline{c}_i$, where:

$$\overline{c}_i = \frac{\sum\limits_{j \neq i} c_j}{n-1} \tag{5}$$

However, the price rule could be:

$$p_i = \min\left(c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n\right) \tag{6}$$

or, alternatively,

$$p_i = \max\left(c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n\right) \tag{7}$$

In the event that agents have different exogenous characteristics, $\mathbf{x}$, Shleifer suggests a regression procedure:

$$p_i = \overline{c}_i + \mathbf{b}'\left(\mathbf{x}_i - \overline{\mathbf{x}}_i\right) \tag{8}$$

Any of these mechanisms will result in each firm choosing to operate at the first best level of average cost[3] , and receiving the same level of economic rent. Bös (1991) on the other hand points out the obstacles to the yardstick competition outcome, which include: comparability issues amongst agents, techniques and search costs, e.g. sampling error, measurement error, and omitted variables, possible collusion amongst agents, lack of commitment by the principal to fulfil the incentive contract after the costs are observed, and restrictions on the size of the penalty that can be imposed on a participating agent who does not reveal a low cost technique of production.

In addition, it appears that despite statements about using efficiency and productivity analysis to develop transparent and widely acceptable incentives, many regulators use primitive methods, small samples and ad hoc adjustments in applying benchmarking to determine the levels of the price control. It is important therefore to understand why this is happening. To accomplish this, it is helpful to have a simple model of regulation in which efficiency and productivity analysis can be embedded, and, to keep the discussion simple, consider restricting the theoretical analysis to linear regulatory contracts. All of the case studies in section 2 of the paper use regulatory contracts which are linear. Three well known and useful linear regulation models are Schmalensee (1989), also used by Joskow (2005), Armstrong et al (1994) and Gasmi et al (1994). These models are very close to each other in structure and results, and that of Armstrong et al (1994) is taken as representative[4] . The basic definitions and results are shown here, with the justification left to the appendix.

The regulator seeks to set a price (or revenue) control by relating the regulated price, P to the observed cost, c, without observing managerial effort, e, which can also be regarded as productive efficiency. P can represent either average or total revenue, and c can represent either average or total cost and a linear model is:

$$P = \overline{P} + (1 - \rho)\, c \tag{9}$$

$\overline{P}$ is the fixed element of the price control, and the parameter ρ is called the incentive power of the control. The higher the value of ρ the weaker the relationship between the price or revenue cap and the firm's observed cost.

In the case of a single regulated firm, the regulator chooses $\left(\overline{P}, \rho\right)$ to minimise the level of the price cap or revenue cap to the firm, taking effort, e, as given by the firm's choice, The firm chooses e to maximise its utility taking $\left(\overline{P}, \rho\right)$ as given by the regulator's choice. Assume the disutility of effort function, known to the regulator, is $\psi(e)$ with derivative equal to marginal disutility of effort:$\psi'(e)$. This enables the regulator to take this function into account when designing a price control. Although the regulator cannot observe effort, he knows that cost, which can be verified, depends on both effort and a random variable, outside the firm's control. Therefore:

$$c = \theta - e \tag{10}$$

The regulation problem is that there is not a perfect correlation between the firm's effort or productive efficiency and its observed cost: $r_{ce} < 1$. However, the parameters of the distribution of the random component in cost are common knowledge:

$$\theta \tilde{}\, d\left(\mu, \sigma^2\right) \tag{11}$$

The firm's profit is revenue[5] minus cost and the disutility cost of effort.

$$\pi = P - c - \psi(e) \tag{12}$$

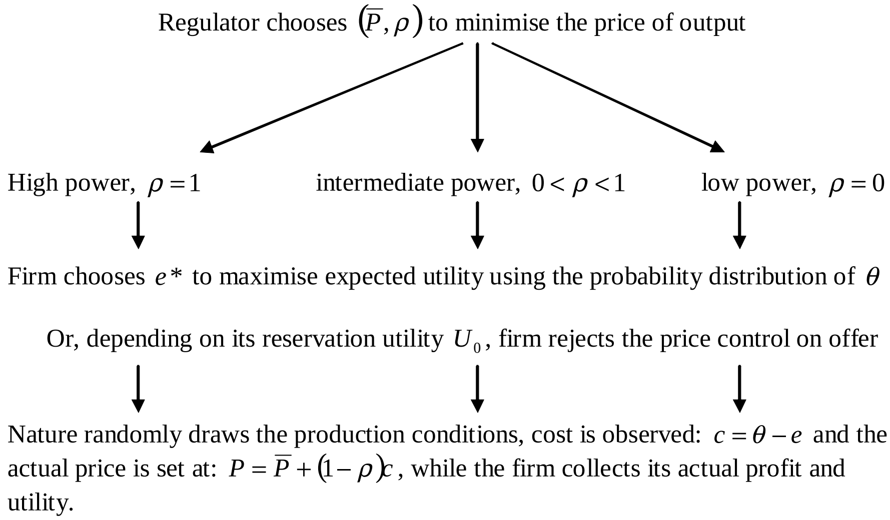

Figure 1 Extensive form of a regulation game

Will the firm accept the contract and participate in the game? This will depend on its reservation utility , which is the amount of utility (in monetary terms) that it could gain in an alternative economic activity. Finally the firm's utility function indicates that it enjoys economic rent but dislikes its variability.

$$U = E(\pi) - \frac{\gamma}{2} var\ \pi \tag{13}$$

The extensive form of the game is illustrated in figure 1.

It can be seen that this is a two stage game, with the regulator choosing the price control parameters first, then the firm choosing the effort level when it knows these parameters and if it accepts the regulatory contract, by operating on the basis of expectations about the random variable that will affect production conditions.

Begin with the firm's sub-game, taking the regulator's choice of $(\overline{P}, \rho)$ as given. The utility maximising firm chooses effort to reflect the incentive power of the regulatory contract by setting the marginal disutility of effort to whatever value ρ of the regulator has chosen.

$$\psi'(e^*) = \rho \tag{14}$$

Substituting $e^* = e^*(\rho)$ into the utility function, the firm's equilibrium expected utility level from participating in the game depends on the regulator's price control formula, the mean and variance (risk) of the random variable representing production conditions, and the manager's risk aversion coefficient,γ.

i.e.:

$$U^* = U^* \left(\overline{P}, \rho; \mu, \sigma^2, \gamma\right) \tag{15}$$

The regulator's subgame has the following solution: choose $\left(\overline{P}, \rho\right)$ given that $\psi' \left(e^*\right) = \rho$, and ensuring that the participation constraint will be satisfied: $U^* - U_0 \geq 0$

The solution is the pair of parameters:

$$\rho^* = \rho^* \left(\sigma^2, \gamma\right) \tag{16}$$

and

$$\overline{P}^* = \overline{P}^* \left(U_0, \rho; \mu\right) = \overline{P}^* \left(U_0; \mu, \sigma^2, \gamma\right) \tag{17}$$

Interpret as follows. The variable part of the price control ρdepends on the coefficient of risk aversion and the amount of risk represented by the variance of the random variable. The fixed part $\overline{P}$depends on three factors: the reservation utility that must be covered to induce participation in the game U_0, the mean of the random variable representing production conditions μ, and the variable incentive power part of the price control ρ. If the principal chooses the value of ρsub-optimally, say by opting for a high-powered contract $\rho = 1$when the agent is risk averse, then all that happens is that the expected utility of the agent will be reduced, $\frac{\partial U^*}{\partial \rho} < 0$, and non-participation, i.e. rejection of the regulator's contract becomes more likely[6] . If the agent does participate nevertheless, effort will be higher than in the case of optimal choice of ρ. In summary the effect of choosing a sub-optimally high powered contract is:

$$\rho' > \rho^* \Longrightarrow U' < U^*, but\ e' > e^* \tag{18}$$

For investor-owned utilities, the share-holders can reject a regulatory contract by selling their equity in the firm to other participants who will take on the regulator's contract at a reduced share price. This possibility is closed off when the regulated firms are local municipal networks as is often the case in European regulatory jurisdictions.

The model can extend to yardstick competition. An example is the regulation of several regional water supply companies, or several regional electricity distribution monopolies.

Armstrong et al (1994) suggest a regulatory price formula that takes the form:

$$P = \overline{P} + (1 - \rho)\, c_i + \rho\, (k c_j) \tag{19}$$

Here the price control for firm i contains a fixed element, plus a weighted average of its own observed marginal cost and a proportion (k) of the observed marginal cost of another comparable firm j. The regulator is engaged in two simultaneous 2-stage games with each company. How does this change the solution?

Begin by thinking of the relationship between the costs of the firms. For two firms, we have:

$$c_i = \theta_i - e_i, i = 1, 2 \tag{20}$$

Consequently it is still the case that there is less than perfect correlation between observed cost and effort or productive efficiency: $r_{ce} < 1$. A simplification is to assume that the variance of the random variable is the same for both firms, and to write: $cov\left(c_i, c_j\right) = r\sigma^2$, where $r$is the correlation coefficient between the random variables representing productive conditions in each of the two firms. Each firm independently sets the marginal disutility of effort equal to the regulator's choice of the cost pass through parameter:

$$\psi'\left(e_i^*\right) = \rho \tag{21}$$

Consequently, the regulator expects each firm to adopt the same effort level or productive efficiency level in equilibrium – unless each has very different preferences about the disutility of effort. The regulator has three variables to consider: k is the proportion of the other company's observed cost which is weighted in with the regulated company's marginal cost; $\overline{P}, \rho$ are the fixed part of the price control and the incentive power of the optimal contract. The solution for $k$is very simple:

$$k = r \tag{22}$$

the proportion of the other firm's cost which is counted is equal to the correlation coefficient between the random variables determining the production conditions. This solution also impacts on the others:

$$\rho^* = \rho^*\left(\sigma^2, \gamma, r\right) \tag{23}$$

and

$$\overline{P}^* = \overline{P}^*\left(U_0, \rho; \mu\right) = \overline{P}^*\left(U_0; \mu, \sigma^2, \gamma, r\right) \tag{24}$$

The major impact lies in the correlation between the random variables representing production conditions. If $r = 0$, then the optimal contract reduces to the single agent case. If $r = 1$, then the optimal contract reduces to:

$$\overline{P} = \overline{P}\left(U_0\right) \tag{25}$$

$$\rho = 1 \tag{26}$$

$$P_i = \overline{P} + c_j \tag{27}$$

The firm has the highest incentive power because $\rho = 1$and it keeps all of the cost savings it makes from reducing its own costs. This is the maximum incentive power even if the firm is risk averse and faces highly variable costs whatever its effort level. It is however fully insured because it can pass on the full level of costs observed from its competitors in the yardstick mechanism, and it receives the same revenue whatever the state of the world, and the level of its own marginal cost[7] .

The decision about the power of the regulatory contract entirely determines the firm's effort in reducing cost. If the regulator chooses a value for the power parameter that does not reflect the firm's preferences, then the achieved level of utility will be lower and the participation constraint will be violated. Setting a contract that is too high powered simply leads to fewer participants accepting the regulatory contract.

In practice however, efficiency and productivity analysis plays a major role in both the single firm and multiple firm case. The model treats the regulator's sub-game as an analytical optimisation exercise, but this does not generalise to reality. In the real world, each regulator needs a numerical algorithm to determine the level of the price or revenue control that meets the regulatory objective of minimising the capture of economic rent from consumers by the regulated firm subject to the financial viability of the participating firms that accept the contract. Since an analytical solution is not available in reality, the numerical algorithm must focus on the participation constraint; for example, in the multiple firm case this is:

$$U^* \left(\overline{P}, \rho, k\right) - U_0 \geq 0 \tag{28}$$

Treating U_0 as the money metric required for financial viability, efficiency and productivity analysis can be used to determine a numerical approximation to the feasible level of $[U^* - U_0]$, i.e. the amount of economic rent or slack that can be feasibly transferred from financially viable regulated firms if each locates on the efficient frontier. This efficiency change is conditional on the power of the regulatory contract that is in place. Regulatory benchmarking is needed, therefore, to determine a feasible numerical solution to the regulator's problem of determining the optimal transfer of economic rent. In summary, efficiency and productivity analysis is a device for capturing economic rent.

When efficiency and productivity analysis is understood in this context, it helps to shed light on the some of the curious economic specifications adopted in real world regulatory cases. Four issues were isolated in section 2 of this paper: methodology, sample size, specification and translation. Since the regulator can use efficiency and productivity analysis for rent capture, he or she will prefer a methodology that sets a very challenging target for the firms to appeal against if the cost-benefit ratio of judicial appeal does not deter them. Consequently, methodologies may favour primitive methods such as COLS. The regulator is aware that he or she is only permitted to capture rent from the current cross-section of firms. Previous observations, as in panel data, or firms in other jurisdictions, as in international samples, offer no candidates for rent transfer, and may only serve to make the current cross-section look as if it has relatively little slack compared to a wide international panel. Regulators therefore may prefer small national cross-section samples to correspond to the small sample open-envelopment methodologies which have also been used. In considering the specification of the TOTEX variable the regulator will operate in terms of the corporate finance model that the capital market uses – consequently this may vary from jurisdiction to jurisdiction. Finally, the regulator will tailor the translation of the rent capture to the capital market response that will apply to the firms under jurisdiction, consequently leading to different forms of glide-path and price control. Therefore viewing efficiency and productivity analysis as a numerical algorithm, which is used to substitute for an analytical solution to devise the level of the price cap that captures economic rent, makes more sense of how it has been applied in the real world than thinking of it only as a form of incentive signal.

Two additional issues further complicate the role of efficiency and productivity analysis in regulation, and these will impact on the amount of economic rent that can be captured in a single frontier efficiency study. These issues are quality of supply and dynamic regulation. Firms may substitute between care for quality and cost reduction, so that the firm must be given a lower powered contract, or additional intermediate power contracts related to quality issues must be available when a high-powered contract applies to the regulated price. Laffont and Tirole (1993) argue that where the regulated firm is forward looking and sensitive to its reputation in repeated supply of service then more high-powered schemes may be used. It is essential therefore to include quality of supply measurement when assessing efficiency and productivity analysis applications to regulation. Repeated regulatory review also raises the key issue of dynamic regulation which does complicate matters significantly.

Regulators and firms are in practice engaged in repeated interactions, with possibly fixed periods, e.g. of five years, between regulatory reviews. The key is the extent to which the regulator can perfectly commit to the regulatory contracts offered in the first period. If this is assumed to be so, then the regulator and firm can sign a long term contract which simply replicates the static incentive compatible mechanisms in each period. However such commitment is unlikely for two reasons according to Laffont and Tirole: most jurisdictions have legal restrictions on the ability of regulators to sign such long term contracts – the political and consumer pressures are likely to be too strong, and the problem of incomplete contingent contracts arises whereby unforeseen events lead to re-contracting. Crew and Kleinfdorfer (2002) have argued strongly that lack of credible commitment undermines the incentive compatible mechanism approach to regulation. In this case the firm may fear, and the regulator favour, a ratchet effect whereby the regulator tightens the constraint on the firm as it builds up experience of the firm's capability.

A general way round this problem of the breakdown of incentive compatible regulation, arising from the imperfect ability of the regulator to commit to a long term contract, is to be unable to use the information from the first period. Regulators find it beneficial to delay making use of the frontier efficiency information learned in the repetitions of the game perhaps by preserving the firm's initial incentives over several periods, or increasing the length of time between reviews.

4. Regulatory Benchmarking with DEA

In a series of papers by Peter Bogetoft and others, see especially Bogetoft (1997) and Agrell, Bogetoft and Tind (2000), the authors explore the relationship between yardstick competition and the use of data envelopment analysis, DEA which has been one of the most widely used methods of comparative efficiency measurement amongst regulated utilities. Among n different utilities the typical firm is observed to have input expenditures of $w'x$ where x is a vector of inputs and w is a vector of input prices, a vector of outputs y and possibly a vector of non-controllable inputs z. This data is verifiable in the sense that the regulator can measure and check the data on outputs, input expenditures and non-controllable inputs. The regulator contracts with the firms at the beginning of the game to pay a revenue cap b according to a formula that depends on the observed costs and outputs of the firms.

The utility (but not the regulator) knows the minimal cost of using current technological possibilities to produce the outputs given the inputs, input prices and non-controllable inputs:

$$C\left(y|z,w\right) = \min_{x}\left\{w'x : x \text{and } z \text{ can make } y\right\} \tag{29}$$

The firm (or the managers) can choose a degree of slack, s, which is also unknown to the regulator, so that the actual cost experienced by the firm is:

$$C\left(y|z,w\right) + s \tag{30}$$

The regulated firm's ex-ante utility is assumed to depend on the difference between (i) its allowed revenue cap b and its actual verified input expenditure $w'x$, plus (ii) a fraction[8] $\varphi$of the difference between the expenditure on inputs and the cost (including slack) of producing its output target:

.

$$U = b - w'x + \varphi \left(w'x - C\left(y|z, w \right) - s \right) \tag{31}$$

where the strict inequalities $0 < \varphi < 1$are satisfied. The slack is consumed by the firm in converting inputs into outputs using the available technological possibilities. The restrictions on the marginal utility of the slack, φ, ensure that at the margin the firm prefers to increase profit rather than to consume slack although both yield positive utility. The regulator is unaware of the minimal cost function, but knows or estimates the firm's marginal utility of slack, and endeavours to minimise the informational rent paid to each regulated firm through the revenue cap.

The result is that generally an optimal (individually rational and incentive compatible) revenue cap contract which will minimise the amount of informational rent to be paid to the firms takes the following form for each firm:

$$b = w'x + \varphi \left(c^* - w'x \right) \tag{32}$$

i.e.

$$b = \varphi c^* + (1 - \varphi)\, w'x \tag{33}$$

where c^* is a "best practice cost norm" or "minimal extrapolation cost standard" set to act as a benchmark for the firm. The benchmark must provide an upper bound on the informational rent paid to the firm so it is essentiak that it at least exceeds the minimal technological cost of production:

$$c^* \geq C\left(y|z, w \right) \tag{34}$$

Consequently it is required that c^* is the least upper bound of the possible values of the cost of production. Without knowing the minimal technological cost of production the regulator has to find the least upper bound of the set which contains this unknown function. The observed input expenditures, outputs and non-controllable inputs of the firms that are subject to the yardstick competition can provide information about this least upper bound. In particular, under assumptions of disposability and convexity of the production possibility set, the DEA efficient cost under constant returns to scale, $C^{DEA-CRS}$ could be a candidate for c^*: However, for incentive reasons the benchmark should exclude the cost and output of the firm in question from the reference set for which the frontier is calculated. In this respect the suggestion replicates the DEA model of Andersen and Petersen (1993) which was initially suggested as a way to rank firms all of which are efficient according to the standard DEA model. In general however, these arguments provide both a model of yardstick competition and an analytical justification for using the DEA frontier efficiency measure.

An issue that dominates the debate on methods of efficiency and productivity analysis in practice is the non-stochastic aspect of DEA. Procedures have been developed to address the problem, notably by Simar (2003), Simar and Wilson (2000) and Banker (1993), but also by Land, Lovell, and Thore (1993), and Lovell (1993), Ruggiero (2004), and Seaver and Triantis (1989). Nevertheless it is difficult to find any major regulatory use of stochastic developments in DEA by regulators, and the fundamental division between DEA and econometric approaches remains to be addressed in the future by many regulators.

5. Regulatory Benchmarking with Parametric SFA

The simple model of the agent's observed marginal or average cost can be expressed in terms of the notation used in econometric models of performance measurement[9] .

$$c = \mu + u + v \tag{35}$$

where the variables are:
μ: the frontier level of cost ($\mu = E(c) = \alpha + \beta x$in a regression framework)
u: the agent's chosen level of effort or inefficiency
v: stochastic error – e.g. measurement error, sampling error, omitted variables
The interpretation of the inefficiency term as a choice variable with a predictable outcome in the principal agent model is recognisable as one of the arguments used by Leibenstein (1966) to explain X-Efficiency, and as the major component of the Farrell (1983) analysis of X-Efficiency.

In performance measurement there is the advantage of observing several agents and the regulator can use their individual costs or outputs as benchmarks. With two comparators, costs are:

$$c_i = \mu_i + u_i + v_i \tag{36}$$

and

$$c_j = \mu_j + u_j + v_j \tag{37}$$

Suppose stochastic factors are absent, and that costs vary only because the different agents operate with different levels of inefficiency.

$$c_i = \mu_i + u_i \tag{38}$$

The correlation between observed cost and inefficiency is exactly equally to 1.

$$r_{cu} = 1 \tag{39}$$

Rasmusen (2001: 171) concludes: "unobservability of effort is not a problem in itself, if the contract can be conditioned on something which is observable and perfectly correlated with effort. The true agency problem occurs when that correlation breaks down"

Consequently, while there is a measurement problem to be resolved when stochastic factors are absent, nevertheless this specification does not address the true problem of modelling the behaviour of agents who are better informed than the principal. It is the presence of other stochastic factors that cause difficulties for the principal or regulator. In Rasmusen's (2001: 171) words: " because of uncertainty about the state of the world, effort does not map cleanly onto the observed output". Adapting the Armstrong et al (1994) model to consider the correlation between observed cost and inefficiency when the observation also contains stochastic error:

$$c_i = \mu_i + u_i + v_i \tag{40}$$

Now the correlation is:

$$r_{cu} = \sqrt{\frac{var\ u}{var\ v + var\ u}} < 1 \tag{41}$$

and

$$r_{cu} \longrightarrow 0\ as\ var\ u \longrightarrow 0 \tag{42}$$

The observation on cost provides a partial but imperfect source of information about inefficiency, and the correlation is weaker the smaller the proportion of cost variation that is accounted for by inefficiency. This is now exactly the nature of the principal-agent problem, and it is exactly captured by the stochastic frontier analysis (SFA) model of performance measurement.

Suppose both agents have accepted the regulator's incentive contract, and that this contract requires that the performance benchmark for one agent is the observed cost of the other. If the agents do not differ strongly in their aversion to effort or increased efficiency, their response to improved incentives will very similar at the margin. Then, it is likely that they will choose very similar levels of effort or efficiency. This time look at the correlation of their observed costs:

$$\begin{aligned} r_{c_i c_j} &= \frac{cov\,(u_i u_j) + cov\,(u_i v_j) + cov\,(v_j u_i) + cov\,(v_i v_j)}{\sqrt{var\ u_i + var\ v_i}\sqrt{var\ u_j + var\ v_j}} \\ &= \frac{cov\,(u_i u_j)}{\sqrt{var\ u_i + var\ v_i}\sqrt{var\ u_j + var\ v_j}} > 0 \end{aligned} \tag{43}$$

This correlation, in a Nash equilibrium of the yardstick competition game is likely to be relatively high, especially if the two firms are subject to similarly distributed stochastic errors. It is still assumed that stochastic errors and inefficiency are independent of each other.

Now consider what the standard models of performance measurement have to say about these same correlation coefficients between observed cost and inefficiency, and between the observed cost of different companies. The answers are not necessarily the same as the game theory predictions.

The COLS model of inefficiency measurement and the DEA model as well, emphasise inefficiency as the main or sole factor explaining performance variability:

$$c_i = \mu_i + u_i \tag{44}$$

so that:

$$r_{cu} = 1 \tag{45}$$

Consequently, we should expect COLS to be an inadequate guide to situations which are best modelled as principal-agent games. On the other hand the SFA model explicitly sets out a composed error model of output or cost. For example in the normal-half normal case

$$c_i = \alpha + \beta x_i + \varepsilon_i \tag{46}$$

$$\varepsilon_i = u_i + v_i \tag{47}$$

$$u_i \tilde{\ } NID^+ \left(0, \sigma_u^2\right), i = 1 \ldots n \tag{48}$$

$$v_i \tilde{\ } NID \left(0, \sigma_v^2\right), i = 1 \ldots n \tag{49}$$

The objective is to estimate the following parameters:

$$\lambda = \frac{\sigma_u}{\sigma_v} \tag{50}$$

or

$$\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} \tag{51}$$

These are exactly the relationships at the core of the game theory model of principal and agent. Although other forms of hypothesis are in fact used, nevertheless, the basic hypothesis testing procedure of the SFA model can be expressed as:

$$H_0 : r_{cu} = 0 \tag{52}$$

against:

$$H_1 : r_{cu} > 0 \tag{53}$$

Rejecting H_0 in favour of H_1 is equivalent to finding significant levels of inefficiency in the performance data. In this context, note that COLS is not a useful substitute if estimation of the SFA model breaks down. SFA may break down for several reasons, especially, for example, small sample size. This leads to failure to reject $H_0 : r_{cu} = 0$. Replacing SFA with COLS corresponds to the strategy of assuming: $r_{cu} = 1$. So far, the SFA model is a better reflection of the principal-agent problem than the COLS model.

However, the SFA model deviates strongly from the game theory model in its prediction about the correlation between the observed cost of different firms. The SFA model, as shown above, includes the assumption that the components of the composed error are independently distributed. This implies zero correlation between both the inefficiency and stochastic error components of different disturbance terms.

$$r_{c_i c_j} = \frac{cov\,(u_i u_j) + cov\,(u_i v_j) + cov\,(v_j u_i) + cov\,(v_i v_j)}{\sqrt{var\ u_i + var\ v_i}\sqrt{var\ u_j + var\ v_j}} = 0 \tag{54}$$

If this were not true then we would have the problem of error terms that were correlated with each other in cross section samples. The SFA model therefore is designed to capture a situation in which agents display different, non-correlated, inefficiency outcomes. This is completely different from the usual principal-agent framework. A tempting supporting argument is that if the different agents have different risk-effort preference structures, then an optimum could be found in which each agent received a different incentive contract designed for his/her preferences, in which case we would expect to observe different, non-correlated efficiency choices. But this argument is faulty, because the nature of the game has changed into one in which agents may have an incentive to falsify the signals about their preferences. It does not provide an escape route from the dilemma that the SFA model is based on an assumption that is not compatible with one prediction of standard principal-agent game theory. A finding of inefficiency in the SFA model may be an indicator that the underlying principal-agent game has not been solved efficiently, or that non-optimal contracting arrangements are being used.

Developing this theme of the relationship between the SFA model of inefficiency and the asymmetric information game, Gagnepain and Ivaldi (2002) investigate the relationship between the specification of the SFA model and the principal agent, PA, framework for regulatory economics. These authors extend the application of the PA model to include both hidden action (moral hazard) and hidden information (adverse selection). They envisage a model identical to the Armstrong et al model except for the addition of an idiosyncratic error term leading to a model in which the composed error term has three components, e.g.:

$$\varepsilon = u + v = \theta - e + v \tag{55}$$

Here v is the usual idiosyncratic error, but their inefficiency term has two components:θ is exogenous technical inefficiency, and $e \geq 0$is the effort productivity exerted by the firm. Their PA model states that the regulated firm's utility is:

$$U = t_0 + \phi \left(R - C\left(e, \theta\right)\right) - \psi\left(e\right) \tag{56}$$

$$\phi \in \{0, 1\} \tag{57}$$

The term t_0 is a fixed sum reflecting the resource transfer to the firm which would be required to ensure that expected revenues covered expected costs. The term $\psi\left(e\right)$is the usual disutility of effort and the term $\phi\left(R - C\left(e, \theta\right)\right)$reflects the amount of the firm's positive or negative profits retained by the firm. The incentive power of the regulatory contract is given by the term ϕ; a zero value indicates a cost plus contract with the regulator bearing all of the risk, while a value of 1 indicates that the firm bears all of the risk. Gagnepain and Ivaldi demonstrate that when this model is used to derive the firm's cost frontier relating cost $C$to explanatory regressors x, the result is [10]:

$$\ln C = \phi\left[\text{terms in } \theta, x, \psi\left(e\right)\right] + \left(1 - \phi\right)\left[\text{terms in } \theta, x\right] = v \tag{58}$$

With a low powered regulatory contract $\phi = 0$the usual classical least squares model can be applied, but with a high powered incentive contract the values of the explanatory variables x, i.e. input prices and outputs, are endogenised. In other words they become correlated with the effort level chosen by the firm. Since this effort level forms part of the inefficiency component of the SFA model, we can summarise by the result:

$$r_{xu} \neq 0 \tag{59}$$

This endogeneity issue could be serious and will lead to inconsistent estimates of the regression coefficients as well as difficulty in identifying the inefficiency arising from low effort. The endogeneity of the inefficiency term and its effect on the regression explanatory variables requires the regulator to think carefully about how the model is estimated. Two procedures are available. Gagnepain and Ivaldi prefer an approach that explicitly derives the appropriate log likelihood function based on particular assumptions about the specification of the effort disutility, and the production function, as well as assumptions about the error components themselves. They derive a Cobb-Douglas based example for estimation[11] . A simpler approach (also noted by Gagnepain and Ivaldi) is to use a fixed effects, FE, panel data model.

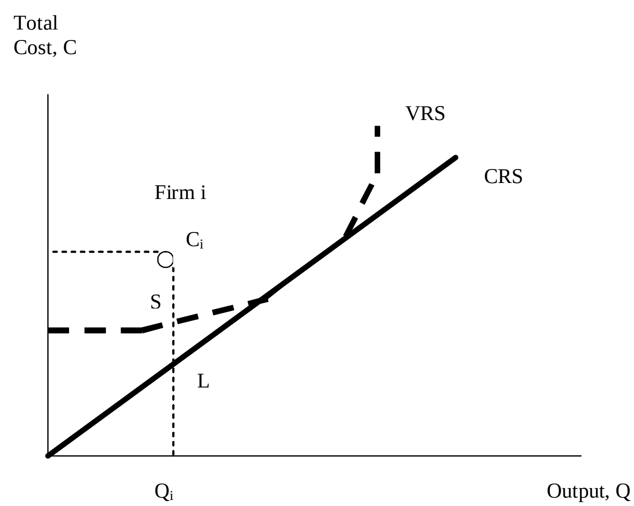

Figure 2 DEA Scale Efficiency

6. Scale and Envelopment Issues in Regulatory Applications

In practical regulation cases, the effect of returns to scale becomes an issue. One source of inefficient performance may be that the regulated firms are constrained by history or regulation to be sub-optimal in size. Regulators may debate whether to treat the scale of firms as outside their control, or to penalise firms for apparently choosing to remain at a sub-optimal size, see Burns et al (2005), and CEPA (2003b). Partly this concern arises because the DEA approach seems to offer the regulator the option of whether or not to include scale efficiency as a target. Figure 2 illustrates an example of a regulatory study that has total cost as the input and seeks to measure technical efficiency relative to output, assuming that all the firms face the same input prices.

VRS and CRS frontiers are shown in figure 2. Take an observed firm with data point: (Q_i, C_i). Its DEA-CRS efficiency is θ^{CRS} = [LQ_i / CQ_i]. Its DEA-VRS efficiency is θ^{VRS} = [SQ_i / CQ_i]. If the firm's target performance is set relative to the CRS frontier, it will be much more demanding than for the VRS frontier, and the implication is that the firm will be able to adjust its size to generate the efficiency improvements that the regulator is looking for. For example, such a requirement was assumed by the UK Postal Regulator in benchmarking the individual centres of the monopoly firm UK Royal Mail. However, the ability to alter scale may be very limited for the firm in practice and therefore the size of the efficiency gain targeted by the regulator may simply not be feasible. Figure 2 illustrates another issue, the assumed orientation of the efficiency analysis. In an input orientation, the firm shown appears to lie in the area of increasing returns to scale, but in an output orientation the firm lies in an area of constant returns to scale. Färe et al (1994: 122-3) discuss this phenomenon which arises because of ambiguity about the meaning of scale efficiency.

Can the same problem be assessed in parametric approaches, i.e. can the regulator distinguish between efficiency relative to different scale assumptions? In the econometric model of efficiency and productivity analysis, the approach has to be different because a regression line is fitted with VRS as the default assumption. Fitting a CRS function to the same data will result in specification bias if CRS is not a valid assumption for this sample of data or this data generating process. Restricting the variables in the fitted VRS cost function will not produce the CRS effect properly. However an alternative approach is the following. The long run total cost curve embodies variable returns to scale, while the short run total cost curve by definition keeps some factors fixed at their current scale level. It is then possible in principle to derive the short run total cost curve by restricting the functional form of the fitted long run cost curve.

Chambers (1988) sets this out as follows, where y is a vector of outputs, x is a vector of inputs, w is a vector of input prices, and C is observed total input expenditure. In the long run, total cost is given by minimisimg input expenditure subject to the technology represented by a parametric production function

$$c\left(w, y\right) = \min_{x} \left(w'x : y \leq f\left(x\right)\right) \tag{60}$$

Now partition the inputs into variable and fixed inputs respectively, $x = \left(x^1 : x^2\right)$.

The short run variable cost is:

$$c^v\left(w^1, y, x^2\right) = \min_{x} \left(w^{1\prime}x^1 : y \leq f\left(x^1; x^2\right)\right) \tag{61}$$

while short run total cost is :

$$c^s\left(w, y, x^2\right) = c^v\left(w^1, y, x^2\right) + w'^2x^2 \tag{62}$$

By consequence of the additional constraint on the cost minimisation:

$$c^v\left(w^1, y, x^2\right) + w'^2x^2 \geq c\left(w, y\right) \tag{63}$$

from which the Le Chatelier - Samuelson- principle allows the derivation of the familiar envelope cost function property:

$$c\left(w, y\right) = \min_{x^2} \left[c^v\left(w^1, y, x^2\right) + w'^2x^2\right] \tag{64}$$

However this property holds at every point on the long run total cost function, not only at minimum efficient scale (or Most Productive Plant Size, MPSS in DEA terms).

Then the corresponding idea in parametric econometric analysis to the ideas of VRS and CRS efficiency in DEA would be:

$$Short\ run\ efficiency : SRE : \frac{\left[c^{v}\left(w^{1}, y, x^{2}\right)+w'^{2} x^{2}\right]}{C} \tag{65}$$

and

$$long\ run\ efficiency : LRE : \frac{c\,(w, y)}{C} \tag{66}$$

With particular parametric functional forms it is possible to derive the expression $\left[c^{v}\left(w^{1}, y, x^{2}\right)+w'^{2} x^{2}\right]$ from knowledge of $c\,(w, y)$ so that econometric estimation can in principle be used to derive these corresponding ideas, if very strong assumptions are made about the specification of the technology. However, the assumptions are likely to be so strong that regulatory challenges are almost certain. A preferable approach to addressing scale issues in parametric approaches is the scale adjusted decomposition of total factor productivity used by Coelli et al (2003), and Saal et al (2004) and based on Orea (2002). However this requires the availability of panel data.

One of the motivating forces in DEA is the aim of showing the producer in the best light. By constructing efficiency measures relative to a piecewise linear representation of the production technology, DEA offers the closet envelopment of the data under different assumption regimes, and this was demonstrated earlier in this paper. In fact Petersen (1990) has argued that closest envelopment is an extremely desirable property of he DEA approach to efficiency measurement. The regulatory purpose is to give the benefit of doubt to the regulated firm, Agrell and Bogetoft (2006)

Is there a corresponding idea in parametric econometric approaches to efficiency? The idea is important in a regulatory context because of the possibility of sub-optimal choice of the power of the regulatory contract which can lead to the non-participation of some firms. An econometric frontier has an entirely different conception, but the idea of looking at its envelopment properties is nevertheless possible, and arose in submissions to the UK water regulator's 2004 review, Weyman-Jones et al (2006). In many European regulatory applications, it has already been shown that small sample and outlier-dominated procedures such as COLS have been used for constructing the efficient frontier. However COLS and similar regression frontiers are point estimators of the frontier, and it is possible to base a closest envelopment parametric frontier on an alternative interval estimation approach. This has a regulatory advantage in two ways. First, it allows

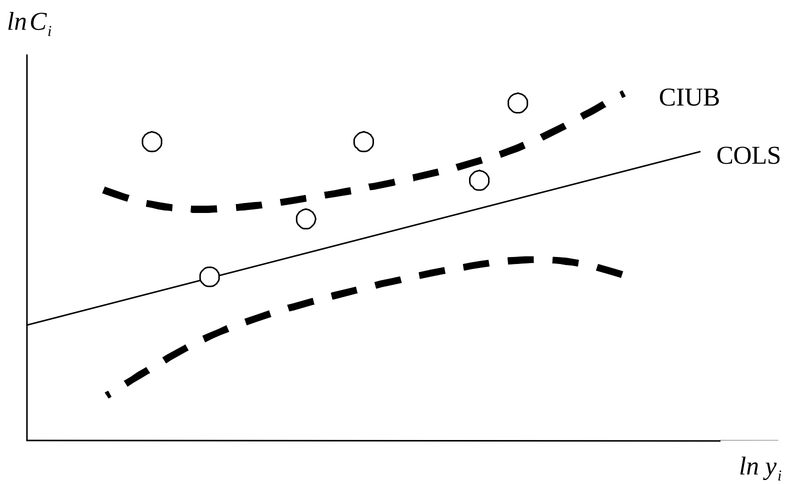

Figure 3 Prediction Interval around the COLS cost frontier

for the possibility that in opting for a high powered contract, the regulator may be extracting a sub-optimal amount of economic rent from risk averse firms, and therefore discouraging participation, and secondly, by incorporating sample second moment properties of the data it takes some account of the risk aversion of the regulated firms. The procedure described also offers an incentive to regulators to improve the quality of the benchmarking process.

The essential idea is illustrated in figure 3.

Take a small sample cross-section case such as those examined in section 2 of the paper. A COLS cost frontier has been constructed passing through the observation with the minimum OLS residual. Since the COLS frontier is a parallel displacement of the fitted OLS regression, a prediction interval can be constructed around the COLS frontier which is simply a parallel displacement of the OLS prediction interval. Instead of the COLS point estimator of the frontier, this prediction interval is an interval estimator of the frontier conditional on a pre-determined confidence level. The interpretation is developed as follows.

Fit the OLS regression:

$$\ln C_i = \widehat{\beta_0} + \widehat{\beta_1} \ln y_i. \tag{67}$$

Define two sets of inefficient observations. Firms which lie above the COLS frontier and are therefore regarded as inefficient are in the set:U^{COLS}

$$U^{COLS} = \left\{ \ln C_i > \ln \widehat{C_i} + \min\left(e_i\right), i = 1 \ldots n \right\} \tag{68}$$

Firms which lie *above* the $100\left(1-\alpha\right)\%$ confidence interval upper bound on the COLS regression are in the set U^{CIUB}

$$U^{CIUB} = \left\{ \begin{array}{c} \ln C_i > \ln \widehat{C_i} + \min\left(e_i\right) + t_{\frac{\alpha}{2}} \left(\sqrt{est\ var\ \left(\ln C_i - \ln \widehat{C_i} \right)} \right), \\ i = 1 \ldots n \end{array} \right\} \tag{69}$$

U^{CIUB} has approximately[12] the property that if samples are repeatedly drawn and $\widehat{\beta_0} + \widehat{\beta_1} \ln y_i + \min\left(e_i\right)$ is calculated each time. then, in $100\left(\frac{\alpha}{2}\right)\%$ of the samples, $\ln C_i$ will be contained in U^{CIUB}, even if efficient. Since $\ln C_i$ is the dependent variable, the variance term can be re-scaled as suggested by Greene (2003:112). In general, there will be considerably fewer observations in U^{CIUB}than in U^{COLS}, and their average efficiency should be considerably higher. This gives an approximate confidence level for stating that an observation is inefficient.

Since the size of the confidence interval is greater the greater is the residual variance, the smaller the sample, and the more extreme is the value of the explanatory variable used to evaluate the distance of the chosen firm from the frontier., this interval estimator of the frontier produces corresponding incentives for the regulator. To capture more economic rent, the regulator is required to increase the sample size, improve the precision of the estimate of the frontier, make allowance for firms distant from the mean, and state an explicit degree of confidence in the determination of whether a firm is inefficient.

If the upper bound of this prediction interval is taken as the frontier, the result is a performance target with the following properties:

1. It is more harsh for the firms (implying higher average inefficiency scores) the lower degree of confidence that is required
2. It is more harsh for the firms (implying higher average inefficiency scores) the smaller the unexplained variation in the residuals.

Take the upper bound of the confidence limits to preserve the incentive to participate in the regulatory game. This has the effect of benefiting the firms if the fit is weak and the confidence level is required to be high. The conclusion of using this interval estimator of the frontier is that, if all of the sample variation is attributed to inefficiency, than it is possible to be confident on the basis of the estimated equation that only the observations in the set are inefficient. The impact on the participation incentives of regulated companies could be substantial. Depending on the confidence level chosen, a restricted number of firms may be found to be inefficient, compared with the very substantial inefficiency scores set for the COLS estimator of the frontier. It remains true that the analysis assumes that all of the residual variation is due to inefficiency, since the SFA model has not been applied. However, the location of the frontier explicitly takes account of the goodness of fit in the sample.

When larger samples are available, then the interval estimator of the frontier can be applied in a stochastic frontier analysis model. Horrace and Schmidt (1996) and Hjalmarsson, Kumbhakar and Heshmati (1996) demonstrate confidence intervals for technical efficiency which can be applied to develop interval estimates of the efficient frontier for regulatory purposes.

7. Efficiency and Productivity Analysis in Governance

Efficiency and productivity analysis is becoming widely used in the public sector, Fox (2002). In this area, the purpose and issues in efficiency and productivity analysis are very different from their application in regulation. Providing incentives to owners of firms does not arise, and managerial reward conditions in the public sector have historically been limited. A notable recent exception in UK has been the linking of medical practitioners' salaries in the UK National Health Service to evidence of quality of patient care, Atkinson (2005).One issue that remains consistent is to explain the divergence in raw performance by different public agencies such as education and health services. For example in many European public services, it is common to develop 'league tables' of performance by schools, colleges, hospitals and emergency services. The purpose of efficiency and productivity analysis is then to reconcile the differences in performance by the identification and inclusion of exogenous operating characteristics until the raw differences are reduced, Vignoles et al (2000).

A major problem for efficiency and productivity analysis in Government and public sector activity is the definition of output. Atkinson (2005) notes that there is a long history in many countries of the public sector's use of the definition: output = input. The Atkinson review is a major attempt by the UK Government to address the issue of finding direct measures of public sector outputs. The European Commission has directed that public sector outputs must be measured directly for inclusion in the national accounts of members of the European Union, but this requires massive research efforts to achieve. Lovell (2002) notes that the reason why output = input is so frequently used is that public sector outputs are usually un-priced, and that quality indicators are undefined. Education and health provision are two areas where performance measurement using frontier methods is widely adopted, Greene (2005), Vignoles et al (2000), and where the issue of output definition has to be widely addressed otherwise inefficiency is always zero, and productivity is by definition always unity. Greene (2005) demonstrates the need for careful definition of health delivery output in a World Health Organization panel dataset by using a composite measure of success in meeting five health service goals. Atkinson (2005) notes that prior to 2004, UK National Health Service output was measured by the number of inpatient and day-cases, accounting for about half of national input expenditure. The Atkinson review recommended the adoption of quality adjusted health years and health outcomes as measures of output. In education, previous UK Government practice used school enrolments as the output measure in the national accounts and Atkinson recommended an immediate change to school attendances. There is a quality adjustment in most education output measures, but it is often based on a single year's attainment at the end of schooling, and Atkinson recommended an ongoing quality attainment adjustment throughout the students' life. In public good provision, the definition of output becomes particularly difficult since a key non-rival and non-excludable aspect of such public services as fire and police protection is the availability of the service on call. It is almost inevitable, as Atkinson recognised, that this output will have an input dimension, and may possibly need to be related to population size, De Borger et al (1994). Although output measurement has been the major issue in applying efficiency and productivity analysis to the governance sector, input measurement also turns out to be a major problem. In computing efficiency and productivity in the public sector, volume measures of input are derived by deflating input expenditures by input price indexes. Efficiency measures become sensitive to the choice of deflator but input price deflators which are relevant specifically to the public sector are particularly difficult to use, and this is especially problematic in the measurement of the price index of capital. Diewert and Lawrence (2002) provide one of the clearest guides to measuring capital costs for efficiency and productivity analysis.

8. Lessons from regulation and governance

It is tempting to close this survey of efficiency and productivity analysis in regulation and governance with a list of major conclusions, but it is clear that a large number of issues are unresolved. The successful application of efficiency and productivity analysis in practice has not kept up with the development of the theoretical and empirical methodology. The overwhelming impression of regulatory and governance case studies is that sample size, variable choice, model specification and choice of methodology has been governed by different objectives from those of the theoretical literature. The nature of regulatory games gives some insight to this. Although the theoretical research interest is in performance measurement for improved efficiency, in actual regulatory applications the purpose has been feasible capture of economic rent from firms under the regulator's jurisdiction. This does form part of the wider picture of efficiency and productivity analysis but leads to different model choice and specifications from those pursued by academic researchers. The picture is not wholly bleak however, because total factor productivity studies, e.g. CEPA (2003b) are beginning to show that even crudely applied efficiency and productivity analysis together with incentive regulation can deliver major gains in performance without endangering firm viability. In the governance sector, efficiency and productivity analysis has helped public service managers to understand better the nature of public service output, so that in several countries the public sector is slowly beginning to move away from using input usage as a measure of performance.

Notes

1. I am grateful to Philip Burns and David Saal for commenting on this paper. Any remaining errors are the responsibility of the author.

2. This will naturally have a UK bias partly because the UK has the one of the longest experiences of both incentive regulation and benchmarking using efficiency and productivity analysis, and partly because of the author's own bias in covering the literature.

3. The first best outcome is characterised by an equilibrium level of average cost at which the net social benefit of investing to reduce average cost by $1 more is just equal to the welfare loss of a $1 rise in price.

4. Schmalensee (1989) is the earliest model, but requires numerical solutions as does Gasmi et al (1994), whereas Armstrong et al (1994) has an analytical solution which aids comparative static analysis.

5. Armstrong et al (1994) fixes the level of output at unity, and assumes a zero elasticity of demand, at least for the range of price variability that is to be modelled. This allows the model to be applied equally to price capping or revenue capping situations.

6. In the sense that with sub-optimal incentive power, the range of values of for which the contract will be rejected expands.

7. Note that each firm is assumed to have the same reservation level of utility determined by the capital market.

8. The notation for the marginal utility of slack differs from that used in these papers to avoid confusion with the notation used by Arsmtrong et al (1994).

9. The analysis in this section uses results from Weyman-Jones et al (2006), and Burns et al (2006)

10. The first order conditions for utility maximisation reflect the usual condition relating marginal disutility of effort to incentive power.

11. Assuming the specification is correct, the maximum likelihood estimator is consistent.
12. This ignores the issue that the minimum residual is endogenous to the procedure for constructing the prediction interval.

9. Appendix

This appendix sets out the mathematical solutions of the two sub-games in Armstrong et al (1994).

The Agent's sub-game

Assume that disutility of effort is expressed in monetary terms as:

$$\psi(e) = \frac{e^2}{2}, \psi'(e) \equiv \frac{d\psi(e)}{de} = e$$

The firm's profit is its allowed revenue minus cost, $c = \theta - e$, and disutility:

$$\pi = P - (\theta - e) - \frac{e^2}{2}$$

The unconstrained maximisation for the risk averse manager of the firm when γ is the coefficient of risk aversion is.

$$\max_e U = E(\pi) - \frac{\gamma}{2} var\ \pi$$

where

$$E(\pi) = \overline{P} + (1-\rho)(\mu - e) - (\mu - e) - \frac{e^2}{2}$$

and

$$var\ \pi = \rho^2 \sigma^2$$

First and second order conditions are:

$$\frac{dU}{de} = -(1-\rho) + 1 = 0; \frac{d^2U}{de^2} = -1 < 0$$

There is a unique global maximum where:

$$e^* = \psi'(e^*) = \rho$$

To calculate the agent/firm's level of utility in the optimising equilibrium, substitute $e^* = \rho$ into :

$$U^* = E(\pi) - \frac{\gamma}{2} var\pi$$

so that

$$U^* = \overline{P} + (1-\rho)(\mu-\rho) - (\mu-\rho) - \frac{\rho^2}{2} - \frac{\gamma}{2}\rho^2\sigma^2$$

or simply:

$$U^* = U^*\left(\overline{P}, \rho; \mu, \sigma^2, \gamma\right)$$

The Principal's Sub-game

Now substitute the optimal level of the agent's effort: $e^* = \psi'(e^*) = \rho$ wherever the unobserved value of effort is to be calculated, and minimise the expected value of the price control, i.e. the firm's regulated revenue:

$$\min_{\overline{P},\rho} E(P) = \overline{P} + (1-\rho)E(c) = \overline{P} + (1-\rho)E(\theta - e^*) = \overline{P} + (1-\rho)(\mu-\rho)$$

This is subject to the constraint that the equilibrium expected utility of the agent/firm is at least as high as its reservation level of utility, U_0 :

$$U^* = \overline{P} - \rho(\mu-\rho) - \frac{\rho^2}{2}\left(1+\gamma\sigma^2\right) \geq U_0$$

Examination of the Kuhn-Tucker conditions rules out corner solutions, since it is assumed that the regulator is interested in an incentive contract $(\rho > 0)$, unwilling to allow the firm surplus rent $(\lambda > 0)$, where λ is the Kuhn-Tucker multiplier, and that the firm's reservation utility exceeds zero, requiring: $\overline{P} > 0$. Consequently the first order conditions for the principal's problem yield

$$\overline{P} = U_0 + \rho(\mu-\rho) + \frac{1}{2}\rho$$

i.e.:

$$\overline{P}^* = \overline{P}^*(U_0, \rho; \mu) = \overline{P}^*\left(U_0; \mu, \sigma^2, \gamma\right)$$

and:

$$\rho = \frac{1}{(1+\gamma\sigma^2)}$$

i.e.:

$$\rho^* = \rho^*\left(\sigma^2, \gamma\right)$$

Yardstick competition model
Each firm's problem is now:

$$\begin{aligned}\max_e U &= \overline{P} + (1-\rho)(\mu_i - e_i) - \rho k (\mu - e_j) - (\theta - e_i) \\ &\quad - \frac{e_i^2}{2} - \frac{\gamma}{2}\left(\rho^2\sigma^2\left(1+k^2-2kr\right)\right)\end{aligned}$$

$$\frac{dU}{de} = (1-\rho) + 1 - e_i = 0 \Longrightarrow e_i = \rho$$

Use this solution to find the level of utility from participating in the game:

$$U^* = \overline{P} - \rho(1-k)(\mu-\rho) - \frac{\rho^2}{2}\left(1+\gamma\sigma^2\left(1+k^2-2kr\right)\right)$$

The regulator's problem is:

$$\begin{aligned}\min_{\overline{P},\rho,k} E(P) &= \overline{P} + (1-\rho)(\mu-\rho) + \rho k(\mu-\rho) \\ s.t. \quad U^* &\geq U_0 \\ \overline{P} &\geq 0, \rho \geq 0, k \geq 0\end{aligned}$$

Solve as before to obtain:

$$k = r$$

$$\overline{P} = U_0 + \rho(1-k)(\mu-\rho) + \frac{1}{2}\rho$$

$$\rho = \frac{1}{(1+\gamma\sigma^2(1-r^2))}$$

The same caveats apply: sub-optimal choice of ρ leads to greater non-participation, but increased effort on the part of firms which do participate.

10. References

Agrell, Per and Peter Bogetoft, (2006) Model Specification, Bundesetzagentur (2006) Incentive Regulation in the German Electricity and Gas Sector – efficiency and reliability to set the yardstick, International Scientific Conference of the Bundesnetzagentur (Federal Network Agency), Bonn

Agrell, Per, Peter Bogetoft and Jorgen Tind (2000) Multi-period DEA incentive regulation in electricity Distribution, North American Productivity Workshop, 2000

Andersen, and Petersen (1993) A Procedure for Ranking Efficient Units in Data Envelopment Analysis, Management Science, 39,1261-4

Armstrong, Mark, Simon Cowan and John Vickers (1994) Regulatory Reform: Economic Analysis and British Experience, Boston, MIT Press, 1994

Atkinson, Tony (2005) Atkinson Review: Measurement of Government Output and Productivity for the National Accounts, London, Palgrave

Banker, Rajiv D. (1993) Maximum Likelihood, Consistency and Data Envelopment Analysis, Management Science, 39, 10, October, 1265-1273

Bogetoft, Peter (1997) DEA-based Yardstick Competition: the optimality of best practice regulation, Annals of Operations Research, 73, 277-98

Bös, D. (1991) Privatization: a theoretical treatment, Oxford: Clarendon Press.

Bundesetzagentur (2006) Incentive Regulation in the German Electricity and Gas Sector – efficiency and reliability to set the yardstick, International Scientific Conference of the Bundesnetzagentur (Federal Network Agency), Bonn

Burns, Philip, Cloda Jenkins and Christoph Riechmann (2005) The role of benchmarking for yardstick competition, Utilities Policy 13, 302-309

Burns, Philip, Cloda Jenkins and Thomas Weyman-Jones (2006) Information Revelation and Incentives in Michael Crew and David Parker (eds) International Handbook of Economic Regulation, Cheltenham, Edward Elgar

Carrington, Roger, Tim Coelli, and Eric Groom (2002) International Benchmarking for Monopoly Price Regulation: the case of Australian Gas Distribution, Journal of Regulatory Economics, 21, 2, March, 191-216

CEPA (2003a) Assessing efficiency for the 2005 distribution price control review, Cambridge Economic Policy Associates for Ofgem (UK Office of Gas and Electricity Markets) London.

CEPA (2003b) Productivity Improvements in Distribution Network Operators, Cambridge Economic Policy Associates for Ofgem (UK Office of Gas and Electricity Markets) London.

Chambers, Robert G., (1988) Applied Production Analysis: a dual approach, Cambridge, Cambridge University Press

Coelli, Tim, Antonio Estache, Sergio Perelman, and Lourdes Trujillo (2003) A Primer on Efficiency Measurement for Utilities and Transport Regulators, Washington, The World Bank, 2003

Crew, Michael and Paul R. Kleindorfer (2002) Regulatory Economics: Twenty Years of Progress, Journal of Regulatory Economics, 21(1) 5-22

De Borger, B., K. Kerstens, W. Moesen, and J. Vanneste (1994), Explaining differences in productive efficiency: an application to Belgian municipalities, Public Choice, 80, 339-58

Diewert, Erwin and Dennis Lawrence (2002) The Deadweight Costs of Capital Taxation in Australia in Kevin J. Fox (ed) Efficiency in the Public Sector, Boston, Kluwer Academic Publishers, 104-167

Färe, Rolf, Shawna Grosskopf and C. A. Knox Lovell, (1994) Production Frontiers, Cambridge, Cambridge University Press.

Filippini, Massimo, Medhi Farsi and Aurelio Fetz (2005) Benchmarking Analysis in Electricity Distribution, Centre for Energy Policy and Economics, Swiss Federal Institutes of Technology, CEPE Working Paper 39, March

Farrell, Joseph (1983) Monopoly Slack and Competitive Rigor: a simple example, Massachusetts Institute of Technology mimeo, February 1983, reprinted in Eric Rasmusen (ed) Readings in Games and Information, Oxford, Blackwells, 2001

Farrell, Michael (1957), The Measurement of Productive Efficiency, Journal of the Royal Statistical Society A, CXX (Part 3) 253-90

Gagnepain, Philippe and Marc Ivaldi (2002) Stochastic Frontiers and Asymmetric Information Models, Journal of Productivity Analysis, 18, 145-159, 2002

Gasmi, F., M. Ivaldi, and J.J. Laffont (1994) Rent Extraction and Incentives for Efficiency in Recent Regulatory Proposals, Journal of Regulatory Economics, 6, 151-76

Greene, William G. (2003) Econometric Analysis, 5th edition, New Jersey, Prentice-Hall

Greene, William G. (2005) Fixed and Random Effects in Stochastic Frontier Models, Journal of Productivity Analysis, 23, 7-32

Hargreaves, John, Matt Parr, Helen Lay, and Melvyn Weeks (2006) Historical Review of Evidence and Ofwat's Approach to Efficiency, Indepen Consulting Limited, London

Hjalmarsson, L., S.C. Kumbhakar and A. Heshmati (1996) DEA, DFA, and SFA: a comparison, Journal of Productivity Analysis, 7, 2/3, July, 303-28

Horrace, William C., and Peter Schmidt (1996) Confidence Statements for Efficiency Estimates from stochastic frontier models, Journal of Productivity Analysis, 7, 2/3, July, 257-82

IPART (1999) Regulation of Electricity Network Service Providers, Incentives and Principles for Regulation, Independent Pricing and Regulatory Tribunal of New South Wales, http://www.ipart.nsw.gov.uk

Jamasb, Tooraj and Michael Pollitt (2003) International benchmarking and regulation: an application to European electricity distribution utilities, Energy Policy, 31, 1609-1622

Joskow, Paul (2005) Incentive Regulation in Theory and Practice: Electricity Distribution and Transmission Networks, National Bureau of Economic Research Conference on Economic Regulation, Washington

Laffont, Jean-Jacques and Jean Tirole (1993) A Theory of Incentives in Procurement and Regulation, Cambridge, Massachusetts, MIT Press

Land, K.C., C.A.K. Lovell, and Sten Thore (1993) Chance Constrained Data Envelopment Analysis, Managerial and Decision Economics, 14, 541-554

Leibenstein, Harvey (1966) Allocative Efficiency vs. X-Efficiency, American Economic Review, 56, 392-415

Lovell, C.A.K. (1993) Production Frontiers and Productive Efficiency in H. Fried, C.A.K. Lovell and H. Schmidt (eds) The Measurement of Productive Efficiency, Oxford, Oxford University Press

Lovell, C.A. Knox (2002) Performance Assessment in the Public Sector, in Kevin J Fox (ed) Efficiency in the Public Sector, Boston, Kluwer Acaemic Publishers

Orea, L. (2002) A Parametric Decomposition of a Generalised Malmquist Productivity Index, Journal of Productivity Analysis, 18, 1, 5-22

Petersen, N.C. (1990) Data Envelopment Analysis on a relaxed set of assumptions Management Science, 36, 305-14

Pollitt, Michael (2005) The role of efficiency estimates in regulatory price reviews: Ofgem's approach to benchmarking electricity networks, Utilities Policy, 13, 279-88

Rasmusen, Eric (2001) Games and Information An Introduction to Game Theory, 3rd edition, Oxford, Blackwells, 2001

Ruggiero, John, (2004) Data Envelopment Analysis with stochastic data, Journal of the Operational Research Society, 55, 2004, 1008-1012

Saal, David, David Parker, and Thomas Weyman-Jones (2004) Determining the contribution of technical, efficiency and scale change to productivity growth in the privatized English and Welsh water and sewerage industry: 1985-2000, Aston Business School Working Paper RP0433 December 2004

Schmalensee, Richard (1989) Good regulatory regimes, RAND Journal of Economics, 20, 3, Autumn, 417-36

Seaver, Bill L., and Konstantinos P. Triantis (1989), The Implications of Using Messy Data to Estimate Production-Frontier-Based Technical Efficiency Measures, Journal of Business and Economic Statistics, 7, 1, 49-59

Shleifer, Andrei (1985) A Theory of Yardstick Competition, Rand Journal of Economics, 16, 3, (Autumn) 319-327

Simar, Leopold (2003) Detecting Outliers in Frontier Models: a simple approach, Journal of Productivity Analysis, 20, 391-424

Simar, Leopold and Paul W. Wilson (2000) Statistical Inference in Nonparametric Models: the state of the art, Journal of Productivity Analysis, 13, 49-78

Vignoles, Anna, Rosalind Levacic, James Walker, Stephen Machin, and David Reynolds (2000) The Relationship between Resource Allocation and Pupil Attainment: a review, Value for Money Unit, London, UK Department of Education and

Skills

Viljainen, Satu, Kaisa Tahvanainen, Jukka Lassila, Samuli Honkapuro and Jarmo Partanen (2004) Regulation of Electricity Distribution Business Nordic Distribution and Asset Management Conference

URL:<http://powersystems.tkk.fi/nodac2004/
papers/nordac2004_honkapuro_et_alpaper.pdf>

Weyman-Jones, Thomas, Julia Boucinha, Celia Godhino, Catarina Feteira Inacio, Pedro Martins, and Jon Latore (2006) Efficiency Analysis and Incentive Regulation in Tim Coelli and Denis Lawrence (eds) Performance Measurement and Regulation of Network Industries, Cheltenham, Edward Elgar