Note for Ofgem on Alternative Methodologies

Professor Andrew Smith, University of Leeds, June 2019

1. Introduction

This note is intended to guide Ofgem in its discussions with companies about the selection of cost efficiency models for use in economic regulation[1]. Its focus is on the potential benefits of adopting alternative techniques to those used at the last price control (which focused on ordinary least squares regression models). The main alternatives discussed are:

- data envelopment analysis (DEA); and
- stochastic frontier analysis (SFA).

All of these methods are used widely in the academic literature but DEA and SFA have had much more limited use in regulatory practice.

2. The motivation for applying different methods

At the last price control Ofgem adopted corrected ordinary least squares regression (COLS) models to set cost allowances. This method permits cost variation between companies to be explained by relevant cost drivers - though in all cases only a single (sometimes composite) cost driver was included in the models. In this way differences between companies can be controlled for prior to making an assessment of relative efficiency. Cost allowances are then based on the predicted cost relationship estimated (representing an average company), with an overlay to represent an efficiency challenge (upper quartile).

There may be multiple motivations for testing alternative methods to COLS. In general, it is good practice to apply the available relevant methods to the data. Different methods typically have advantages and disadvantages and hence it is useful to test alternatives. Where similar methods produce similar results this can enhance confidence in the overall findings. It is also important for regulators to justify their choice of method from the range of alternatives.

We discuss each of the main alternative methods in turn below. The general advantages and disadvantages of the methods are considered, but applying these

[1] I acknowledge the helpful comments of Dr Phill Wheat, Institute for Transport Studies, University of Leeds, on an earlier draft of this note.

arguments to the case of benchmarking the gas distribution businesses is the main aim of this note.

3. Data envelopment analysis (DEA)

The workings of DEA

Data envelopment analysis is a so-called non-parametric efficiency measurement method. The use of the term "non-parametric" can be interpreted to mean that we are not using statistical techniques and we are not estimating specific elasticity parameters for the relationship between costs and the drivers of costs as in the standard ordinary least squares (OLS) regression technique typically used by UK economic regulators. Since we do not estimate elasticities, we are also therefore not making (and cannot make) an assessment as to whether the cost driver is a statistically significant factor in explaining cost variation.

DEA works by specifying a set of inputs and outputs. Inputs may be in physical terms, for example, number of staff, or they may be measured in monetary values (costs). Outputs could be measures such as the composite scale variable used in previous Ofgem gas distribution cost models, or customer numbers for example (in principle monetary measures of outputs could be included).

Once the inputs and outputs are specified, DEA works by mathematical optimisation, selecting weights for the inputs and outputs (potentially different weights for each firm) to compute an efficiency score for each firm. The weights selected for each firm are those which give the most favourable view of the firm's relative efficiency score.

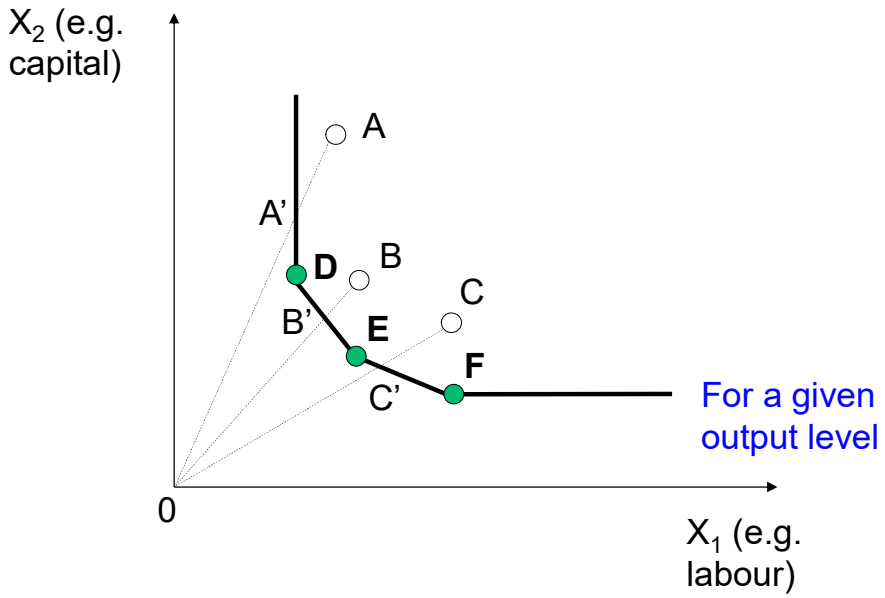
Although different weights may be chosen for each firm's calculation, within the process of computation of a firm's efficiency score, the same weights are applied for all firms. Thus whilst the programme may give low or even zero weights to some inputs and outputs in the process of deriving the efficiency score for, say, Firm A, those same weights are applied to the other firms when computing Firm A's efficiency score.

The above implies that, although DEA selects the weights on the inputs and outputs for each firm to give each firm its highest possible efficiency score - in order to achieve a good score, the firm still has to be better than other firms when using those same weights. Put another way, for a firm to be efficient it still has to find some dimension in which it is as good or better than other firms (i.e. it needs to have one output / input ratio where it performs well). A challenge, however, in small datasets, is that if a firm has an extreme position in terms of its usage of a particular input, or its production of a particular output, it can obtain a high score because there are no other firms producing in that region of the production set.

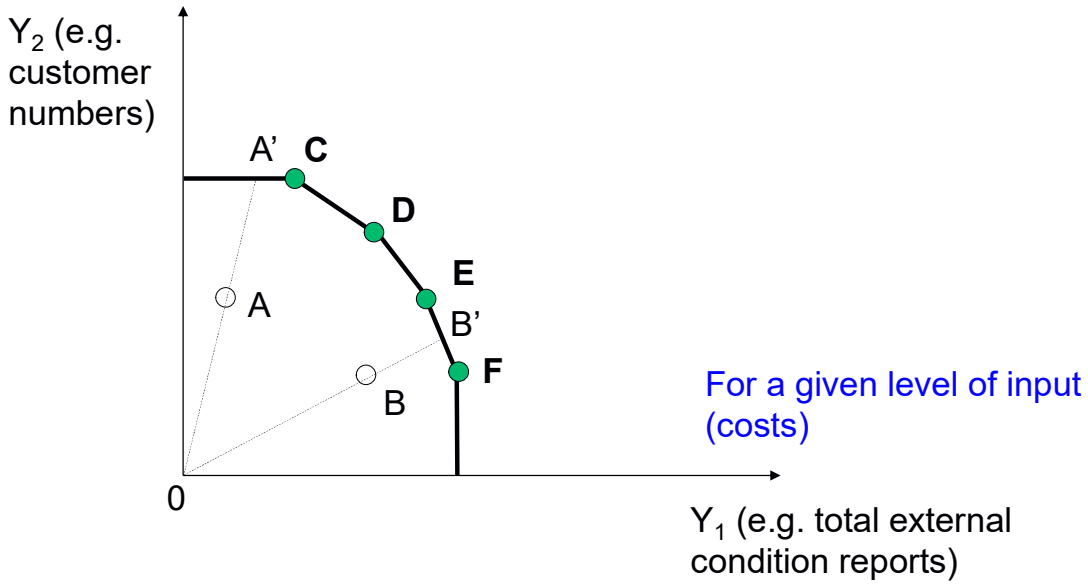
The same process then continues for each of the other firms and each time a (potentially different) set of weights is chosen for each firm in calculating the score for that firm.

DEA can be operated based on the assumption of constant returns to scale. If this assumption is deemed unsuitable for the technology, a variable returns to scale option (VRS) exists which decomposes inefficiency into technical inefficiency and scale inefficiency (the latter reflects the fact that the firm may be too small or too large and this may generally be seen to be outside the control of the firm). DEA also may have an input or output orientation – the former envisages a situation where firms are assumed to be able to contract inputs (costs) for a given set of output requirements; whereas the latter assumes that firms have a fixed budget and are aiming to produce the maximum levels of output from that given budget. The input orientation would typically be assumed for regulated industries in the UK.

The following diagrams illustrates the DEA method for the constant returns to scale case. The first considers an input oriented case where there is a single output, say customer numbers, and two inputs (say capital and labour). The second shows an output oriented case where there is a single input (costs) and multiple outputs (say customer numbers and total external condition reports[2]). In the first case firms D, E and F are technically efficient in that they use the minimum amount of inputs to produce a given output level (the model controls for the fact that firms have different output levels). Firms A, B and C are inefficient and their efficiency is measured by the ratio 0A'/0A, 0B'/0B, and 0C'/0C, respectively. With information on input prices it is also possible to determine the degree of allocative efficiency (that is, whether firms are using the optimal combination of inputs, given input prices).



[2] These two drivers make up the Emergency Composite Scale Variable (CSV).



In the second case firms C, D, E and F are technically efficient in that they are able to produce the maximum levels of outputs from a fixed level of inputs (the model controls for the fact that firms have different input levels). Firms A and B are inefficient and their efficiency is measured by the ratio 0A/0A' and 0B/0B' respectively.

Both diagrams illustrate that it is possible for firms to be viewed as efficient if they have a very high ratio of one of the outputs to one of the inputs. This may be deemed undesirable, particularly where it is considered that firms should be performing well across all of the required output measures as discussed further below.

Advantages and disadvantages of DEA in general

DEA is a candidate model for efficiency analysis. One of its advantages is that it shows each firm in the best possible light and this can be attractive in gaining support amongst companies. This feature can enhance confidence in the sense that firms consider they are not being disadvantaged and are being afforded the most favourable view of their relative efficiency.

A further advantage is that it does not require the assumption of a particular functional form for the technology – this being an issue with the econometric approach in economic regulation, where there is a debate about the use of the Cobb-Douglas and the translog, which can be a challenge to resolve. That said, as will be clear from the diagram above, DEA does in any case effectively impose a shape on the technological relationship between inputs and outputs which is less transparent from the results of the DEA programme and in principle is open to challenge.

An associated disadvantage is that DEA does not produce an estimate of the impact of the cost drivers on costs – i.e. it does not produce coefficients (elasticities) as in the econometric approach. This means it is not possible for companies and others to get a sense of whether the model is reasonable or not. It can be seen to be a black box and therefore not meeting the important criterion in regulation of transparency.

As noted both constant returns to scale (CRS) and variable returns to scale (VRS) versions of DEA exist. However, it can be the case when sample sizes are small that VRS DEA produces results that imply all firms are on the frontier; so in practice CRS and VRS results may be used and interpreted alongside each other. This potentially implies that the DEA literature is more ready to accept the assumption of constant returns to scale (CRS) – that is, an elasticity of costs with respect to output of unity – even though that would not normally be accepted as reasonable in econometric modelling without further testing.

Further, giving a different set of weights for different firms (and potentially with zero weights on some inputs and outputs[3]) may be considered problematic in an economic regulatory framework. As noted above, this would indicate firms with a particularly high ratio of one output to one input are efficient, which may not be credible in a regulatory context.

One development in the DEA literature has been the use of a multi-stage process, where a set of core inputs and outputs are included in the first stage, and the resulting scores are then regressed on a set of exogenous factors deemed to impact on efficiency. An issue with this approach is that it is not always clear what variables should be included in each stage[4]. Secondly, this approach becomes in any case a (partly) econometric approach, thus raising questions therefore about the value of the approach compared to the standard econometric cost function approach.

Finally, DEA assumes that all deviations from the frontier are due to inefficiency – so there is no allowance for the influence that random factors may impact on the data in particular years for particular firms. That said, the corrected ordinary least squares (COLS) method also suffers from this problem, and it could be dealt with by aiming away from the frontier and instead targeting companies with achieving upper quartile performance as UK regulators typically do in their application of COLS[5]. It should be noted that DEA has not been widely used in economic regulation in the UK.

---

[3] It should be noted that the standard, expected shape of input-output relationships in production economics does mean that firms may choose different input or output mixes and thus operate at the extremes of a given input-output space. In this sense, according zero weights to different inputs and outputs could be seen to be reasonable; however, given the constraints faced by regulated firms, it may be less so given the expectation that firms should be producing minimum levels of key outputs.

[4] Whilst the literature suggests that the first stage variables define the frontier whilst the second stage variables cover "environmental" factors impacting on a company's position from the frontier, this still leaves room for ambiguity. For example, whether a variable such as density is an exogenous factor explaining efficiency, or whether it really represents a normalised measure of output (e.g. customer numbers per size of network).

[5] Stochastic DEA approaches exist but they are not concerned with disentangling noise from inefficiency, but rather with the problem that what is desired is the measurement of inefficiency relative to the population of available comparators, not just the sample at hand.

Advantages and disadvantages of DEA for Ofgem

One factor for consideration in determining the best approach for Ofgem is that the models used by Ofgem comprise a single input (a measure of cost) and a single output (e.g. a composite scale variable (CSV)). In this case the DEA frontier (under the standard constant returns to scale model) would be a straight line and the approach would be identical to computing unit costs (e.g. totex per CSV) and ranking them. The only purpose of using DEA in this context would be to apply the variable returns to scale (VRS) approach to DEA which would make a correction for the possible existence of increasing or decreasing returns to scale (though such an approach may fail to differentiate well between firms as noted above).

However, we consider that to deal with the issue of VRS it would be more transparent to estimate a cost function as that would give an estimate of the elasticity of costs with respect to CSV – thus estimating the degree of returns to scale directly. Indeed a key advantage of the econometric approach as compared to DEA is that it gives information about the impact that cost drivers have on costs, and enables debate about whether those relationships are reasonable (including on whether a translog functional form is reasonable or not). Whilst that process can be challenging it has the benefit of being transparent. DEA in contrast can be seen as something of a black-box in this regard and arguably hides the issues faced by researchers seeking to implement the econometric approach, rather than avoiding them.

One possible use of DEA might be to study the impact of including all the elements of, say, the totex CSV measure directly as separate outputs in the DEA model. However, such an approach would most likely lead to zero weights being given to some outputs and, given the sample size, mean that most or all firms would be deemed 100% efficient. As noted above, different firms would be assigned different weights. Such an approach does not seem to be a promising area of further analysis therefore. It would also differ fundamentally from the approach implicit within Ofgem's current methodology, where the composite scale measures for example utilise a common set of weights for all firms.

Overall, whilst DEA could be used as a supporting method, its usefulness may be limited. Nevertheless it is good practice to test alternative techniques and explain the reasons for selecting (or not) the different methods in the final efficiency determination.

4. Stochastic Frontier Analysis (SFA)

Motivation for SFA

The previous section discussed DEA which, as noted, is a non-parametric approach.

We now move to consider parametric techniques in which statistical techniques are used to estimate parameters (elasticities) that indicate the relationship between costs and relevant cost drivers. Statistical techniques are widely used in UK economic regulation, most notably in the form of the corrected ordinary least squares (COLS) method.

The standard ordinary least squares (OLS) cost function may be written as follows:

$$C_{it} = f(Y_{it}, W_{it}, N_{it}, Q_{it}, \tau_t; \beta) + v_{it} \quad (1)$$

for firm *i* in time period *t* and where:

- C_{it} – is the relevant measure of cost
- Y_{it} – represents output (there could be multiple outputs, and also other drivers of cost)
- W_{it} - represents input prices (though these may not always appear if costs have been adjusted prior to estimation)
- N_{it} - exogenous network characteristic variables
- Q_{it} – quality measures
- τ_t - represent time variables capturing technical change or other unobserved effects influencing costs over time[6]
- β – are parameters (elasticities) to be estimated
- v_{it} = random noise, capturing for example (symmetrical) random shocks to the production process and measurement errors.

Under this approach there is assumed to be no inefficiency in the model. All firms are assumed efficient and any deviations from the estimated regression line are assumed to be due to (symmetrical) random error as noted above.

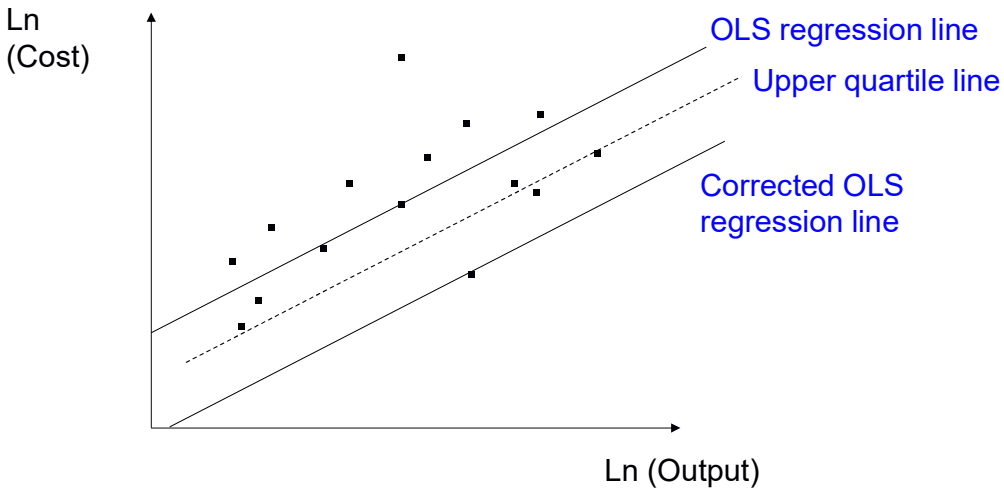
In real applications of course it might reasonably be assumed that some firms may be more efficient than others. Thus the aim of the regression analysis should be to estimate a cost frontier (best practice) representing the most efficient firms. Such an approach would then see deviations from the frontier for two reasons: (1) random error as in the standard cost function; and (2) variation in efficiency performance.

In most UK regulatory applications estimation of the cost frontier proceeds in two steps. First, ordinary least squares (OLS) regression is applied. This method fits a line through the data that minimises the sum of the squared residuals; and will have some firms above the line and some below the line. This is problematic given that

[6] Factors of relevance here could include, for example input price trends over time (influencing all firms in a similar way) if these have not been dealt with via prior adjustments to the costs.

the aim is to estimate a frontier, since some firms are below the cost line which, if it was truly a frontier, would not be possible. This problem is then resolved in the next stage through shifting the OLS regression line downwards (retaining the same slope) until it passes through the firm with the largest negative residual. In this way all firms are either on or above the frontier (corrected OLS line). This can be shown in Figure 1 below.

Figure 1



A key problem with the COLS approach is that it assumes that all deviation from the shifted frontier line represent inefficiency. Further, the position of the frontier is highly influenced by the firm with the largest negative residual, which could result from a one-off event that reduces costs for that firm in a particular year that is not indicative of what can realistically be achieved for firms in general. Partly for this reason regulators may make an upper quartile adjustment, which draws the shifted frontier based on upper quartile performance (lower quartile of the residuals in the case of a cost function) – see Figure 1. The upper quartile line is believed to represent a more reasonable assessment of what efficient firms could deliver.

The motivation for applying SFA compared to OLS is to explicitly allow for inefficiency in the model, which OLS does not. As compared to the corrected version of OLS (COLS), SFA permits a decomposition of the residual (gap between the firm and the regression line) between inefficiency and random noise (whereas COLS assumes all deviations from the shifted regression line to be attributed to inefficiency). The motivation for SFA as compared to the application of an upper quartile adjustment or the application of some other regulatory judgement is less clear however as will be discussed below.

The workings of SFA

There is an extensive literature on SFA and a wide range of alternative models. We start with the basic pooled SFA model as this is most directly comparable to the COLS modelling framework. As noted above the SFA method explicitly allows for inefficiency in the modelling framework. The model is identical to the OLS cost function model set out above, except that an additional term is added, u_{it}, representing inefficiency.

$$C_{it} = f\ (Y_{it},\ W_{it},\ N_{it},\ Q_{it},\ \tau_t;\ \beta)\ +\ v_{it}\ +\ u_{it} \qquad (2)$$

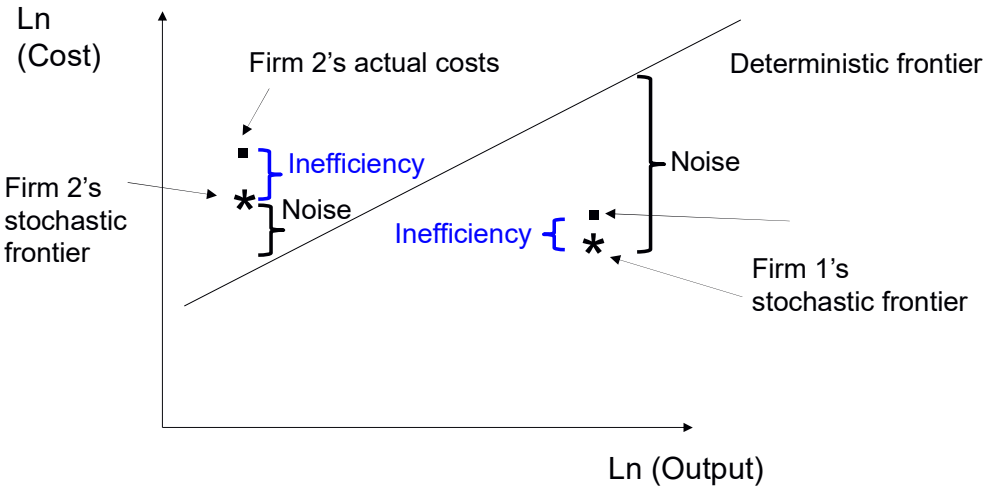
Deterministic frontier Noise Inefficiency

Stochastic frontier

Model estimation proceeds by maximum likelihood estimation, with assumptions being made about the distributions of the random noise term (v_{it}) and inefficiency terms (u_{it}). The random noise term, as is assumed in OLS when making inference about the model parameters, is assumed to be normally distributed with mean zero (and can take positive or negative values). The standard assumption for the inefficiency term is that it has the shape of a normal distribution, but is assumed to be one-sided, so that inefficiency is either zero or positive; hence this distribution is referred to as a half normal. A key assumption is that the noise and inefficiency terms are uncorrelated the explanatory variables.

Thus the model estimates a “deterministic frontier” based on estimating the constant term and the parameters of the explanatory variables – and then a random noise term is added, thus giving a stochastic frontier that reflects the fact that there will be random factors impacting on costs that are nothing to do with inefficiency. The u_{it} term is an estimate of the inefficiency of the firm relative to this stochastic frontier. In this way the overall error term in the model is de-composed into two parts: random noise (u_{it}) and inefficiency (v_{it}). The stochastic frontier model can also be viewed graphically as in Figure 2.

Figure 2



Advantages and disadvantages of SFA for Ofgem

Whilst the academic literature would tend to support the use of SFA techniques over the COLS approach, there are a number of caveats to this statement that apply particularly in the sphere of economic regulation.

First of all, it is recognised in the academic literature that, under the assumption that inefficiency and random noise are not correlated with the explanatory variables (the assumption made in SFA), OLS yields unbiased[7] and consistent[8] estimates of the parameters on the explanatory variables in the model (the slope of the regression line). What is at issue is the estimate of the constant term (intercept) of the model, which impacts on the position of the regression line, and also the related question of the decomposition of the error term – these being required to get estimates of relative efficiency.

[7] This is a finite sample property – the estimated values of the parameters from repeated samples, on average, equal the true population value.
[8] This is a desirable large sample property and means that the parameter estimate approaches the true population value as the sample size increases.

Second, it is also recognised that the pooled SFA model, as applied to panel data, requires assumptions on the noise and inefficiency term that are essentially arbitrary. Third, under the standard distributional assumption noted above (the half-normal model) the SFA technique enables a decomposition of the overall error term into noise and inefficiency, but even after this decomposition, the resulting firm efficiency rankings are the same as those that would result from simply ranking firms based on the overall residuals (as occurs in the COLS approach).

Therefore, whilst the model in a sense decomposes noise and inefficiency, to do so it requires distributional assumptions and also it does not overturn the rankings implied by the overall error term. Thus there is a genuine question as to whether the decomposition achieved by the SFA model is superior (or not) to the application of some form of regulatory judgement, for example through an upper quartile adjustment. A benefit of SFA comes through a potential improvement in the "efficiency" (precision) of the parameter estimates in the model (assuming the distributional assumptions are correct). There is however a general concern over invoking distributional assumptions (which can be seen to be arbitrary) when estimating the parameters themselves.

Particularly given the relatively small samples that Ofgem and other regulators have to deal with, and also to meet the criterion of transparency, OLS could be seen as a preferred approach to estimating the slope parameters (as these will be both unbiased and consistent); leaving the determination of the shift of the function to regulatory judgement. COLS is more transparent in that it is simpler, and also because the upper quartile adjustment is more immediately transparent than the process by which SFA disentangles random noise from inefficiency via the distributional assumptions used.

Notwithstanding the above points, however, we consider that SFA approaches are candidates for inclusion within Ofgem's modelling framework. There exists a range of models from pooled SFA approaches to panel data methods that assume time invariant inefficiency, time varying approaches (structured and unstructured variation), and methods that also seek to decompose the error term into four components[9]. The models that permit structured variation (as opposed to random) in inefficiency over time may also be used to study convergence (or not) in firm scores over time. Some of these models do not require distributional assumptions, though most do, but may require other strong assumptions (e.g. that inefficiency does not vary over time[10]).

Given the relatively long panel available for analysis, we consider that some of these more advanced models could be tried by Ofgem, alongside the relatively simple

[9] See Kumbhakar et. al. (2014): Technical efficiency in competing panel data models: a study of Norwegian grain farming, Journal of Productivity Analysis, 2014, vol. 41, issue 2, 321-337.

[10] This would then simplify to the standard, random effects approach.

pooled SFA approach. These approaches could be compared against the OLS model with COLS-type upper quartile adjustments. Overall though, in this context, and also bearing in mind the need to have transparent models that can readily be communicated and implemented, it may be unlikely that SFA will make an important contribution in this work.

5. Summary and concluding remarks

To conclude, it is good practice for regulators to consider and test alternative approaches. DEA and SFA are therefore candidate approaches that should be explored. By testing a range of alternative approaches the robustness of the overall methodology and results can be enhanced. Depending on the results it may be possible to select a clearly preferred approach for use in setting targets, or alternatively different approaches may be triangulated. It is also possible that alternative approaches are shown not to be useful for inclusion in the suite of models, but nevertheless even in this scenario it is good practice to have tested these models and documented the reason for their exclusion from the analysis.

In terms of SFA starting with a simple pooled SFA approach that is directly analogous to COLS would seem to be the best approach, before exploring more advanced panel techniques. The relatively long time series available could support structured time varying efficiency models that permit the study of convergence (or not) over time[11].

Since in Ofgem's case a single output is used, DEA becomes essentially a comparison on unit costs, though a variable returns to scale version could be adopted to at least address possible economies of scale. DEA might also be used to look at the question of weighting of output measures within a CSV, though with the caveats noted above, and also noting that statistical techniques for assessing that question might be more appropriate.

Given the relatively small sample size available to Ofgem it is far from certain that alternative approaches will yield new insights; also taking into account the complexity / unfamiliarity of both DEA and SFA relative to COLS, and associated transparency

[11] For example, Battese, G.E. and Coelli, T.J. (1992), 'Frontier Production Functions and the Efficiencies of Indian Farms Using Panel Data from ICRISAT's Village Level Studies', *Journal of Quantitative Economics*, vol. 5, pp. 327-348; Cornwell, C., Schmidt, P. and Sickles, R.C. (1990), 'Production Frontiers With Cross-Sectional And Time-Series Variation in Efficiency Levels', *Journal of Econometrics*, vol. 46, pp 185-200; Cuesta, R. A., 2000. A Production Model with Firm-Specific Temporal Variation in Technical Inefficiency: With Application to Spanish Dairy Farms. Journal of Productivity Analysis, 13, 139-158.

issues. It may be that COLS – combined with regulatory judgement (e.g. through the use of upper quartile adjustments) – is a more transparent means of disentangling random noise and inefficiency than SFA.

Finally, the transparency of econometric (COLS and SFA) approaches - in terms of the ability to interrogate the elasticities resulting from estimation and question whether they make sense from an engineering perspective – is an important advantage of econometric approaches over DEA.