Note for Ofgem on Diagnostic Tests in Efficiency Benchmarking Studies

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. Its focus is on the appropriate diagnostic tests to apply in judging whether a model is robust and in comparing and ranking alternative models for use in a price control setting.

2. Prior theoretical rationale

The choice of an appropriate econometric model should start from a theoretical or engineering / business perspective. That is, there should be some prior reason for including a particular variable as a factor explaining cost. This guards against the possibility of "data mining", whereby we are merely picking up spurious relationships between variables.

In the cost modelling literature, based on economic theory, a cost function is defined to be a function of outputs and the prices of inputs (e.g. the wage rate). However, this function can be interpreted more widely as a function of cost drivers, and these may usefully be broken down into the following categories:

- A measure of scale (e.g. length of network).
- A measure of density (e.g. number of customers per length of network); through these first two measures the concept of output is thus separated into a scale and a density measure.
- Exogenous "network characteristic" variables which may reflect, for example, topography or measures of the capability of the network in some sense.
- Quality measures (these could be asset condition, performance measures or even customer satisfaction). Such measures may be deemed problematic because they are under the control of the firm and thus could impact on the robustness of the model in statistical terms, as well as creating incentive issues.
- Depending on the precise cost measure used, it may be appropriate to include so called "intermediate outputs" – e.g. number of new connections or volume of infrastructure renewed.
- Time trend variables (or year dummy variables), capturing technical change or real price effects not captured elsewhere in the model.

The above can be represented in the following equation:

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

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

- Y_{it} - output measures (scale and density); potentially "intermediate outputs" could be included[1]
- W_{it} - input prices
- 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[2]
- β - parameters to be estimated
- v_{it} = random error, intended to capture random factors that might influence costs for a given firm in a particular year (this element may also include inefficiency – see separate note on approaches to modelling[3]).

An important task as part of the modelling process then is to select appropriate variables from (potentially) each of these categories, whilst bearing in mind that when the sample size is relatively small, as in the case of gas distribution, there will be limits to the number of variables that may be included in the model.

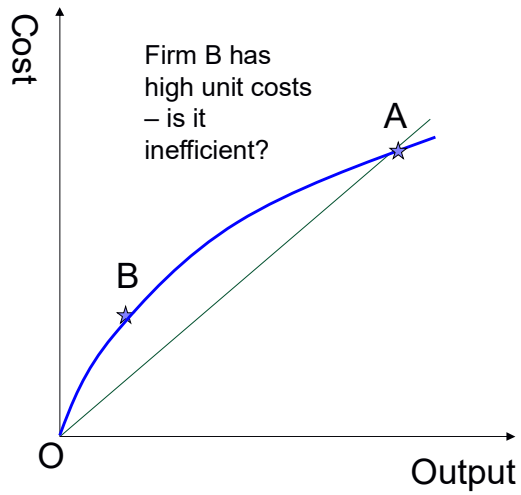
It is worth noting that it is the likely existence of economies of scale (and / or density) in network industries that motivates the use of the econometric approach, as compared to simply using unit cost measures, which implies constant returns to scale (see Figure 1).

[1] For example, length or network replaced or number of new connections (where the cost measure includes capital elements that may be lumpy over time).

[2] 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 or through the inclusion of input prices.

[3] Note for Ofgem on Alternative Methodologies, Andrew Smith, June 2019.

Figure 1



Based on a simple unit cost analysis, firm A would be deemed inefficient as it has higher unit costs than firm B (the straight line from the origin, through point B, measures firm B's unit costs, and using that as a benchmark for firm A implies constant returns to scale). However, if the underlying technology is subject to increasing returns to scale (falling unit costs as output rises) – as implied by the curved line through points A and B), then it becomes apparent that firm A is actually efficient. Its apparent inefficiency resulted from the mistaken assumption of constant returns to scale. As noted below, the econometric model reveals information about the degree of returns to scale based on the data, and it is possible to test whether the constant returns to scale assumption is valid or not.

Whilst it may not be possible to include all types of variables in the model because of lack of data and also problems of multi-collinearity (see below), the likely presence of economies of scale / density means that at the very least some measure of scale is needed, and arguably measures that permit the distinction between scale and density to be isolated. Some measures (either pre- or post-modelling) should be included to capture input price variations. Some form of time related variables (trend or dummies) should ideally be included or at least tested for as it would be expected, for example, that there should be frontier shift over time.

In evaluating the model, it is important to assess whether the modelled relationships between costs and the explanatory different variables are seen to be reasonable compared to business / engineering expectations both in terms of the sign (positive or negative) and size of the estimated coefficients on each variable. A simplified version of the standard double-log, Cobb-Douglas cost function (single output) can be written as follows:

$$\text{Ln } C_{it} = a_0 + a_1 \text{ Ln } Y + v_{it} \quad (2)$$

where the estimated coefficient on each variable (a_1 in the above function) is an elasticity and therefore gives us as view as to what will happen to costs when that variable changes by say 1%. For example, if a_1=1 then a 1% rise in output, Y, would cause costs to rise by 1%, implying constant returns to scale. A coefficient greater than zero but less than (or greater than) one implies increasing (decreasing) returns to scale.

It should be noted that a key advantage of the econometric approach compared to other methods, such as Data Envelopment Analysis (DEA), is that it permits the industry to assess whether the estimated coefficients (elasticities) on each variable are reasonable – which thus then increases the confidence in the use of the model for regulatory purposes.

It should also be noted, however, that one purpose of econometric cost modelling can be to yield new and sometimes challenging information about what drives cost and to what extent. Thus it should be borne in mind that sometimes econometric models will challenge conventional wisdom in this respect. Further, the model is only an approximation to reality, and it is possible that some variables will partly pick up the impact of other variables that are omitted (perhaps because of lack of data).

Finally, consistency with policy is another factor in making model selection decisions. One example might be a case where the inclusion of a measure of quality in the cost function creates an incentive that conflicts or overlaps with other regulatory tools to incentivise quality. This situation could then justify dropping the quality variable from the model, even though there appears to have a theoretical rationale for its inclusion. This issue will be discussed further below.

3. Transparency

For a model to be used in economic regulation it is important that it is clearly explained and can be interpreted by the companies and other stakeholders. These criteria could tend therefore to suggest that it is beneficial to avoid complex econometric techniques and also to choose a parsimonious model. However, there are other considerations and in some cases more complex techniques and specifications may be necessary – and companies can (and do) use their own consultants to run and test these models if needed.

The rationale for selecting the final model or models (as compared to the alternatives) should also be clear. The models should ideally be replicable by the companies, which means that data should be made available and that the methods used should be capable of being implemented in standard econometric packages in general.

4. Statistical robustness of models - overview

The statistical robustness of a model could be defined as covering three broad areas (see also CEPA, 2018[4]; the author of this note was involved in advising on the CEPA (2018) document and therefore the guidance in this note is broadly in line with that of the latter):

- The robustness of the model to appropriate statistical tests
- The stability of the model to changes in, for example, the data sample or precise model specification
- The ability of the model to explain the existing data and to forecast future costs.

It is good practice to set out the criteria that will be used to select models before the model selection process starts.

5. Appropriate statistical tests

The econometric academic literature (including standard econometric textbooks) set out the diagnostic tests that should be carried out to determine the robustness of a model. Here we focus on ordinary least squares (OLS) regression[5] which is the most widely used econometric technique in economic regulation. It is also worth pointing out that "failure" of some tests is more serious than others. The key tests are set out below.

Statistical significance of the coefficients (elasticities).

This test is asking whether we can be confident that there is a relationship between the explanatory variable and cost – or more formally can we (statistically) reject the proposition that there is no relationship (i.e. that the coefficient is zero). Ideally we would only want to include variables where the estimated coefficient (which for the standard double-log model is an elasticity) is statistically different from zero. It is possible to estimate (for example) a positive coefficient, but not to be confident (from a statistical point of view) that the coefficient is different from zero.

A key caveat to this is where several explanatory variables are included in the model together and these variables are correlated – see discussion of multi-collinearity below.

Statistical significance is generally measured at the 1%, 5%, or 10% levels, though there is no set rule on which is the most appropriate. In the OLS model a standard t

[4] PR19 ECONOMETRIC BENCHMARKING MODELS: OFWAT, CEPA (March 2018).
[5] Though we discuss appropriate testing with respect to panel data approaches such as fixed and random effects.

test is used to establish statistical significance. Typically, t-tests greater than around a value of 2 indicate statistical significance at the 5% level.

Statistical significance is also often expressed in terms of p-values, which has a useful interpretation. For example, if we observe a coefficient on an explanatory variable of 0.80, combined with a p-value of 0.01 (1% significance), this can be interpreted as saying that if the true coefficient is actually zero (that is, there is no relationship between cost and the explanatory variable), then there would only be a 1% probability of observing a value of 0.80 (or greater) in a random sample taken from the population. Since we do observe 0.80, and such an outcome would be highly unlikely if the true coefficient was zero, this gives us reasonable comfort that the true value is not zero.

Applying the same argument to the 10% level of statistical significance, if instead the p-value was 0.10 (10% significance), now there is a one in ten chance of observing a coefficient of 0.80 or higher even when the true coefficient is zero. Thus we still have reasonable comfort that that the true value is not zero, but less so than if the p-value is 0.01 (1% significance).

It should be noted that the issue of statistical significance is concerned with whether we think the coefficient is different from zero or not. Establishing this finding may be a rather low hurdle to overcome for a variable such as a composite scale measure, which surely will have a positive coefficient that is likely not to be very close to zero. What may be more important is whether the coefficient is plausible in terms of its size, which is also related to whether we think we have constant, increasing or constant returns to scale. It may therefore be more relevant to be comparing the estimated coefficient to unity rather than zero. Ultimately what is at issue is the precision with which we estimate the coefficients. A positive and precisely estimated coefficient of 0.80 may well be statistically significant when compared to zero, but it may not necessarily be statistically distinguishable from unity.

Standard statistical outputs therefore typically produce ranges or intervals (at the 5% level) to give an indication of how precisely the coefficients are estimated. Thus our estimated coefficient of 0.80 may have a range of 0.7 to 0.9. Provided the range is plausible we can have greater confidence than in a model that produces a much wider range (such as 0.1 to 1.8, where such low and high values might be seen as implausible).

Overall, statistical significance is an important criterion for assessing the robustness of a model as it gives us confidence that the variables included in the model have a real role to play in explaining costs, and also helps us understand how precise our estimates are (say compared to some standard, such as unity in assessing returns to scale). In a model where there is only one variable, lack of statistical significance would typically be a good reason to discard the model.

Multi-collinearity

As noted above it is often the case in economic regulation that variables are correlated (move together). An example might be that a regional wage variable could be correlated with density, for example because wages are higher in London. Multi-collinearity is not serious in one sense – namely it does not cause the estimated coefficients to be biased. They remain unbiased. An unbiased estimator is one that is "right on average". That is, if we were able to take repeated samples from the population then we would get a range of different estimated coefficients but on average these would be equal to the true value.

Thus, even in the presence of multi-collinearity the estimates remain unbiased. However, the standard errors estimated by the model (which affect whether we find variables to be statistically significant – see above) are inflated, thus potentially giving the impression that a variable is not statistically significant. It can also mean the estimated elasticity in the particular sample used may not be "very good" – i.e. it may appear implausible. Of course the sample at hand – i.e. the dataset available to the regulator is the only one available, so this can be a limitation.

Ultimately there is a choice in model selection in this respect. One approach is to leave multiple variables in the model on the grounds that there is a theoretical rationale for the inclusion of each variable and we know that possible multi-collinearity will impact on the individual coefficients. Or, as an alternative, drop one of the insignificant variables from the model to create a simpler model with fewer variables and where all variables are significant. The latter approach may certainly appear advantageous from a presentational perspective, but there is a judgement to be made here.

A risk of excluding a variable is that it will cause bias in some of the other included variables and that its impact will also move into the residual part of the model that is used to compute relative efficiency scores.

In a regulated context the judgement on inclusion / exclusion is finely balanced. It could be argued that a parsimonious model with all variables statistically significant is attractive and can readily be explained to stakeholders. Models with a higher number of variables, where some are statistically insignificant, and with unexpected sizes (or even signs) is much harder to explain. However, provided that there is a clear theoretical / operational / engineering reason for a variable's inclusion, and provided the impact of including the additional variable on cost allowances is clearly explained and understood, then models with more variables can reasonably be selected. In some cases they may well be preferred by some companies who consider that a simpler model does not adequately reflect their circumstances.

One possible area of exploration that Ofgem might consider is whether it is appropriate to include separate elements of Composite Scale Variables (CSVs) in models in place of a single CSV measure. It is likely that some variables would be

statistically insignificant in such an analysis and the above discussion would be relevant as to whether or not to drop some components of the CSV.

It should be noted that multi-collinearity is not a problem that may be tested for explicitly (though the use of simple correlation analysis between variables may be useful, and variable inflation factors (VIFs) can give some indication[6]) – and it will almost always be present to some extent in all models. The only solution to the problem ultimately is to obtain more data; or to make a judgement on whether to include or exclude one of the variables affected.

Omitted variable bias

As noted above, the exclusion of a variable from the model may cause omitted variable bias. An estimated coefficient on an explanatory factor will be biased if there is correlation between the error term and the explanatory variable. This could occur if, because of multi-collinearity between two variables, one of the variables is dropped from the model. In that case the coefficient on the included variable would be a biased estimate of the true value. More generally we know that there will always be some omitted variables, though we cannot be sure as how far they impact on the estimated coefficients in the model.

Of course, practitioners can assess whether they believe the coefficients on the included variables are reasonable or not. Further, if a decision has been taken to exclude a correlated variable because of a perceived multi-collinearity problem (see above) then it can be argued that the variable is now picking up multiple effects and the coefficient can then be interpreted accordingly (i.e. assuming that wage variation is not adequately corrected for in some other way, it may be argued that a density coefficient is also picking up regional wage differentials since where density is high, say in London, wages are also high).

Regulators will always have to deal with limited datasets. They will always face a situation where there are other variables they would like to have but do not have data for. Ultimately there is a judgement to be made as to whether the model is reasonable in any given context. Where there is a decision as to whether to include or exclude a correlated variable this requires judgement to trade off the advantages and disadvantages of dropping a variable. This judgement also applies even if variables do not exhibit high correlations with other variables.

The principal advantage is that the model will be simpler, and all the coefficients will be statistically significant. The disadvantage may be that exclusion of a correlated variable impacts the coefficients on the included variable in an adverse way (makes them appear counter-intuitive) and / or moves the impact of the variable into the error

[6] Variable inflation factors are intended to give an indication of the degree of multi-collinearity and as a general rule VIFs above 10 are considered to indicate high levels, though care is required in interpreting VIFs. As noted above, judgement is still required in deciding what to do in situations where multi-collinearity is deemed to be a problem (whether to drop or retain the variable).

term and thus impacts on the efficiency calculations. Of course a key criterion for the inclusion for a variable in the first place is that it should have a theoretical / operational / engineering reason for inclusion. As noted above under the related discussion of multi-collinearity, a careful balance will be needed on a case by case basis and there can be no a priori fixed statement in favour of simpler versus more complex model specifications.

The RESET test

This test is often referred to in economic regulation by consultants, and the CMA used it in its review of Bristol Water in 2015. Ofwat has also used it in its model selection for PR2019. It is sometimes referred to as a general test for omitted variables. However, it is actually a test for whether there is some non-linear relationship in the model that is not been captured. In the cost modelling literature this is normally dealt with directly by considering a translog specification which captures these non-linearities directly.

A translog model explicitly seeks to incorporate squared and interaction terms for the purpose of approximating complex technologies where, for example, the degree of returns to scale may vary with firm size (note that the scale elasticity is constant in the standard double-log, Cobb-Douglas model, and this assumption is relaxed in the translog case). A simple example, for a single output cost function is as follows:

$$\text{Ln } C_{it} = a_0 + a_1 \text{Ln } Y_{it} + a_2 (\text{Ln } Y_{it})^2 + v_{it} \qquad (3)$$

which is a generalisation of the simple Cobb-Douglas model set out earlier. For that reason, it seems more reasonable to directly test a translog versus the simpler Cobb-Douglas double-log model. Beyond that, it is not clear what else can reasonably be done if a model fails a RESET test and having tested the translog form it would seem overly cautious to reject simply based on the RESET test alone. It should be noted that this test is not widely used in the academic cost modelling literature. That said, given that this test has been used in the regulatory literature, it could be used to distinguish between two otherwise similar models.

Functional form

As noted above a key factor for model selection is the choice of an appropriate functional form (or shape) for the cost function. Linear models are highly limiting because they impose a constant marginal cost for all output levels. Typically, the academic literature uses double-log specifications which are more flexible. As noted above, there is then a choice between the general and highly flexible functional form, the translog, and the more restrictive but parsimonious Cobb-Douglas form. In particular, the translog permits the extent of returns to scale to vary with firm size,

which is potentially an attractive property as increasing returns may become exhausted at a certain firm size.

Since the Cobb-Douglas model is nested with the translog model, the restrictions implied by the Cobb-Douglas can be tested statistically. However, a further consideration is whether the coefficients on the squared and interaction terms in the translog model are intuitively plausible. If they are not, and given the need to communicate and explain the model in a regulatory setting, this could be one factor for regulators to consider in addition to statistical testing. On the other hand, a researcher may prefer to accept that the translog is seeking to approximate a complex technology and therefore accept that this may result in some inability to interpret all of the coefficients perfectly. The latter approach is perhaps easier to adopt in an academic study though than in a regulatory context.

Given the relatively small sample sizes available to Ofgem it is likely that complex translog forms may be hard to justify; however, testing of such forms is justified to examine how returns to scale for example may vary across the sample.

Normality

The breakdown of this assumption does not affect the properties of OLS estimators themselves. They remain best linear unbiased estimators. The impact of non-normality only has implications for the ability to use finite sample inference – that is, making judgements about the statistical significance of the parameters in finite samples. However, whatever the distribution of the error term (whether normally distributed or not), as the sample size increases, the sampling distributions have been shown to be approximately normally distributed. This means we can apply standard inference, though based on asymptotic (large sample) approximations.

Thus the breakdown of normality would not generally be seen to be a problem in most applications, though it could cause issues for very small sample sizes. However, the models used by Ofgem incorporate several firms over several years and thus we would not expect a breakdown of normality to be a serious problem.

Correlation / heteroscedasticity

With respect to heteroscedasticity / autocorrelation, again violations of the assumptions in OLS impact only on the standard errors (that is the standard errors are biased and this impacts on our ability to gauge whether a variable is statistically significant or not) and do not cause the estimates themselves to be biased. The standard response to this potential issue is therefore to use robust standard errors when making assessment of statistical significance.

Testing for panel effects

Given that our dataset comprises observations on multiple firms over several time periods it is a valid question to consider whether models that explicitly recognise the panel structure of the data might be valid alternatives to OLS (the latter simply pools the data and treats all observations as independent).

It is worth noting that the primary issue here concerns the estimates of the parameters in the cost function rather than the approach to estimating inefficiency. The literature (see Schmidt and Sickles, 1984) identifies panel data methods (fixed and random effects) as a means of estimating efficiency by re-interpreting the time invariant random or fixed effect as a measure of inefficiency (benchmarking all firms against the firm with the lowest fixed or random effect in a cost function).

This seems to imply that the random effect method is also a different approach to measuring efficiency – that is, it produces a single efficiency score for each firm over the whole time period that does not vary over time, whereas COLS produces a different efficiency score for each firm in each time period.

However, typically economic regulators use both OLS and random effects models in much the same way to produce efficiency scores (i.e. based on averages over the whole period, rather than year by year scores; for example, Ofwat). Thus the issue simply concerns the difference between the parameter estimates produced by OLS or random (or fixed) effects rather than any fundamental difference in the way the residual and thus inefficiency is treated.

A key determinant with respect to the choice of fixed versus random effects (and in turn OLS) is whether the regressors are correlated with the firm-specific effects[7] (that is, time-invariant effects that are specific to each firm). OLS and random effects both produce consistent estimators of the parameters under the assumption of no correlation between regressors and the firm effects. If there is correlation then the parameter estimates would be biased and inconsistent, but this would affect both OLS and random effects. A Hausman or Wu test may be used to test for correlation[8].

If the assumption of no correlation cannot be rejected via a Hausman or Wu test then OLS, random effects and fixed effects all produce consistent estimates. However, in this case there may be advantages to using random effects because of the "efficiency" benefits in respect of estimation of the parameters (parameters estimated more precisely). The choice between random effects and OLS may also then be judged based on a Breusch-Pagan LM test for random effects, which is essentially testing whether the incorporation of firm-specific random effects adds anything to the model as compared to using OLS.

[7] For the avoidance of doubt, these are firm-specific effects that are distinct from time-fixed effects (year dummy variables) referred to in section 2.
[8] See LIMDEP 11 Econometric Modelling Guide, pages E-434 to E-436.

Where the assumption of no correlation is rejected via a Hausman or Wu test, as a starting point the fixed effects model would be preferred as it produces unbiased estimates of the parameters whether correlation exists or not (which is not the case for OLS or random effects as noted).

However, a key problem in practical applications within the cost efficiency literature in general, and specifically within regulatory applications, is the fact that although fixed effects models produce unbiased and consistent estimates irrespective of whether there is correlation or not between the effects and the regressors, the results for a particular sample may be highly implausible in terms of the size, sign and significance of the parameter estimates. Ultimately there is a trade-off between potential bias and "efficiency" (of the parameter estimates) in comparing fixed and random effects in empirical applications and it may be that random effects (and OLS) produce more plausible estimates in such cases.

Thus, irrespective of the outcome of the Hausman or Wu tests with regard to the non-correlation assumption, fixed effects results may simply be implausible, and typically the choice will be between OLS and random effects in practice. That said, a rejection of the non-correlation assumption should not be ignored altogether and may provoke a re-consideration of the variables to be included in the model (this then relates to the earlier discussion as to which variables to include in the model and which to drop).

As noted, the key choice will usually be between OLS and random effects. Whilst the Breusch-Pagan LM testing might generally indicate a preference for random effects in data with a panel structure, this is not the only consideration and a wider evaluation of the results in terms of the plausibility of the parameter estimates is appropriate. There are also benefits to using OLS in terms of transparency and ease of estimation and its known finite sample properties.

Overall, in regulatory applications it seems likely that the choice will be between OLS and random effects. Appropriate tests exist that can guide the choice of model selected, but the final choice may depend on a wider evaluation of the performance of the model and the plausibility of the estimates.

Endogeneity

In regression analysis the explanatory variables are assumed to be exogenously given and not under the control of the firm. However, this assumption may not hold for some variables, for example, measures of quality. This introduces a possible source of bias since, for example, factors that are omitted from the model (and which are therefore part of the error term) may be correlated with both costs and quality. This might result if managerial decisions (related to their efficiency performance) for example, impact on both costs and the level of quality they deliver.

In general, this problem is complex to address and it is also a question of degree. For example, wage variables are typically included in cost models in the academic literature, even though they may be partly under the influence of the firm.

Consideration of this issue is needed on a case by case basis, and also alongside considerations of how far including a particular variable (such as quality) in the cost model might conflict with other regulatory incentives and targets to improve performance.

6. Model stability

CEPA (2018) includes a detailed section on model stability that the author of this note contributed to. Here we therefore make a number of observations, without repeating the core material there.

In assessing the robustness of a model it is pertinent to ask how the results change with small changes to the sample (dropping firms and years) and to the model assumptions (estimation method). It is also important to ask whether the parameter estimates are stable over time or whether there is a structural break caused by a fundamental change in the industry (for which a Chow test may be used for example).

A key challenge in this area for Ofgem however is that results are likely to change if dropping a single firm from a sample of eight firms, or say one year from a five year panel. In such a case it seems beneficial to work with the larger sample, unless there are very strong reasons for believing that the omitted firm or year are highly unrepresentative. Omitting a firm is highly problematic in regulatory practice of course.

A further, fundamental challenge in regulatory benchmarking is that where a firm has particularly high costs compared to the predictions of the model this could either be because the model specification does not accurately capture the production technology for that firm, or simply because the firm has high costs owing to inefficiency. Attempts can be made in the modelling process to take account of relevant factors, but in some cases this could imply greater complexity which can create additional problems. A pragmatic solution can be to deal with the problem through a combination of special factor adjustments and the use of upper quartile adjustments, whilst keeping all firms within the dataset used for estimation.

Thus exploration of stability is important but it has to be borne in mind the relatively small number of firms available for analysis when deciding the implications of the analysis.

7. Model fit and forecasting

In OLS models the R-squared statistic is the standard measure of model fit. Overall, we are interested in how well the model fits the data – i.e. how well are we explaining the dependent variable. It should be noted that Ofgem typically uses double-log models and R-squared measures are typically higher as a result. In assessing whether the R-squared measure is "good" then it is important to compare against similar double-log cost models.

There can be a danger of overly focussing on this measure by adding more and more variables to the model and thus it is important to have a clear rationale for the inclusion of additional variables, and also to consider the statistical significance of variables added. An adjusted R-squared measure is available that includes a penalty for adding extra variables partly to address this problem.

If we are able to explain much of the variation in the dependent variable we can be reasonably confident in the model and in its ability to predict future years. Provided the explanatory variables in future years are not a long way outside the sample used to estimate the model and provided underlying relationships in the cost function have not changed markedly, we would expect a model with high fit to be good for prediction of the future. Indeed, the rationale for using historic data to generate a model to set cost allowances implies the believe that the past data and cost relationships are a reasonable reflection of the future.

One possibility however, and linked to the earlier discussion, is that a model with many variables, but with a high degree of multi-collinearity, could be unstable and lead to poor prediction. This can be one argument in selecting more parsimonious specifications – however, if the variables that are moving together continue to move together in the future data this should be less of a problem.

8. Summary and concluding remarks

CEPA (2018) summarises the diagnostic tests and indicates a level of priority for each. As noted above the author was involved in production of that report and therefore I do not repeat the ranking there. The comments below are therefore broadly in line with CEPA (2018).

The focus of model development in my view should be, as a starting point, on ensuring high quality and comparable data between firms and over time, and on the development of a strong understanding of the theoretical / engineering / operational rationale for the inclusion of variables in the function. Related to that is the importance of having a transparent process, which also may imply the selection of models that are relatively simple and easy to replicate and communicate. This does not mean that more complex models should be ignored, but rather that they are tested and evaluated, taking into account transparency considerations as well as other factors.

Statistical testing is also important. It is clear from the above sections in this report that there are certain tests that are typically used to evaluate models and it would be desirable if all of these tests were passed. However, in many cases as noted above the consequences of failure of certain tests are relatively small (e.g. normality). In some cases there may be an obvious solution to the problem (e.g. the use of robust standard errors to deal with correlation / heteroscedasticity in the residuals).

In other cases, such as the RESET test, some form of non-linearity may be indicated, but having tried alternative functional forms such as the translog, it is not clear what further can be done. In such circumstances the RESET test would not appear to have much weight, but it could be used to differentiate between otherwise similar models perhaps.

On the related question of translog versus simpler functional forms, a mix of statistical and other considerations may well be required. In particular, a translog model may be preferred statistically, but rejected for other reasons, such as the failure to explain the non-linearity in the model. Transparency (simplicity) could also be another consideration. Whilst in an academic context a researcher may be prepared to accept a translog model that is hard to interpret, on the grounds that it is trying to approximate a complex technology – this may be hard to support in a regulatory context.

Achieving statistical significance of parameter estimates in the model would generally seem to be of high importance as a criteria, but this criterion also needs to be assessed alongside other factors such as the theoretical rationale for inclusion of a variable, and also the challenges of dealing with multi-collinearity, whilst seeking to avoid omitted variable bias. In a single variable model though it would seem highly important for that variable to be statistically significant. Even in that case, however, where the coefficient estimated in a double log model is statistically different from zero, but not statistically different from unity, a simple unit cost model / comparison could be deemed appropriate.

Conformity of coefficient estimates to theoretical / managerial / engineering understanding is important; and certainly it would not seem sensible to include a variable without good reason. However, the challenges of empirical work could mean that coefficient estimates will not always conform to prior expectations and indeed one of the purposes of econometric estimation can be to create new evidence on the impact of variables on costs where none exists. Such estimates could challenge conventional wisdom.

Goodness of fit is also of high importance for regulators as this indicates the extent to which the model is able to explain the data - which in turn has implications for its use in setting future cost allowances. The R-squared statistic is therefore an important indicator (potentially adjusted to take account of the number of variables in the model) and can be evaluated compared to similar double-log models elsewhere in the academic and regulatory literature.

The distinction between random effects and OLS can be tested and in general there could be an argument to say that a model which recognises the panel structure of the data (random effects) may be preferred to one that does not. That said, there are also arguments for falling back on the familiarity and finite sample properties of OLS. Further, given the way that regulators use average efficiency scores to set the efficiency challenge, there is in practice no real difference between OLS and random effects except in terms of estimating the parameters (elasticities) of the model. Whilst tests for correlation between the effects and the regressors should be conducted, given the available data it may be that random effects or OLS is in any case preferable to fixed effects results in practice.

A valid concern in general is the extent to which technologies change over time and therefore testing for stability of the parameters over time is a useful test. More generally, at least in principle, testing the sensitivity of the model to changes in the sample (removing years or firms) can enhance understanding and could indicate the relative instability of some specifications compared to others. However, it also needs to be noted that the sample size that Ofgem has to work with is relatively small. Finally, endogeneity should be considered, but a case-by-case approach is likely to be needed, as this issue is a question of degree and most models will suffer from the problem to some extent. The issue should be considered alongside other issues, and may be particularly important if the inclusion of a variable creates a conflict with other parts of the price control (e.g. with measures of quality).

Overall it is clear that a range of criteria exist, and perhaps greater weight can be attached to some criteria than others. However, often finely balanced decisions are needed. Ultimately it is important that the reasons for model selection decisions are clearly discussed and communicated, with opportunity for comment and input from stakeholders.