# Note for Ofgem on the computation of CSV weights

Professor Andrew Smith, University of Leeds, January 2020

## 1. Introduction

The current approach used by Ofgem for setting cost allowances utilises composite scale variables (CSVs) that capture a range of different output measures within a single variable. The composite measures are computed as multiplicative composites, with the individual output measures raised to the power of a set of weights that sum to unity. A simplified example, using just two components of the Totex CSV, is as follows:

$$CSV = (CONN^{0.37}).(REPEX^{0.63}) \quad (1)$$

where CONN is the connections workload and REPEX is the REPEX workload. Ofgem uses cost shares – that is industry expenditure on the relevant output / activity, as a proportion of total industry expenditure – as the weights. The weights used here (0.37 and 0.63) are indicative, for illustrative purposes.

The idea behind the CSV measure is that the different components may change, possibly to different degrees, and there is a need to weight the changes in output to derive an overall measure of output change. A relationship can then be established between costs (say Totex) and output (as measured by CSV) through a regression model; with the ultimate aim of estimating relative efficiency and setting cost allowances.

As noted, the current approach is to compute a CSV measure using weights based on the industry's expenditure on each of the components; and then include that CSV measure as a single variable in an econometric model. For the Totex CSV the balancing element of cost is represented by MEAV (modern equivalent asset valuation). The alternative approach would be to use an econometric model that includes the different components of CSV separately. This latter model estimates the cost impact of each of the components directly.

The remainder of this note explains the two approaches, and the relationship between them, in more detail, before making some suggestions in terms of next steps.

## 2. The assumptions of a logarithmic cost model

Logarithmic models are widely used by economic regulators within the cost assessment process in the UK and elsewhere. Consider a standard logarithmic model where the natural logarithm of Totex (Ln Totex) is regressed on the natural logarithms of the CSV components individually (using the simplified case identified above):

$$Ln\ Totex = a + b\ Ln\ CONN + c\ Ln\ REPEX + error \qquad (2)$$

where b and c are the elasticities of Totex with respect to CONN and REPEX respectively. An elasticity represents the sensitivity of costs with respect to small changes in the explanatory variables in the model. For example, an elasticity of 0.5 would indicate that a 1% growth in the relevant explanatory variable would lead to a 0.5% increase in cost.

The elasticities can also be expressed as follows:

$$b = MC_{CONN}.\frac{Q_{CONN}}{Totex}$$
$$c = MC_{REPEX}.\frac{Q_{REPEX}}{Totex} \qquad (3)$$

where MC denotes marginal cost of each of the components, and Q represents the volumes relating to the different components of the index.

By estimating a model such as Equation (2), the elasticities are freely estimated, being revealed by what the data is telling us. In turn, the elasticities reveal information about the underlying marginal costs of the different outputs / activities, based on the data, as shown in Equation (3). In order to draw a comparison with Ofgem's current approach, it is useful to express the relative elasticities from Equation (3) as follows:

$$\frac{c}{b} = \frac{MC_{REPEX}}{MC_{CONN}}.\frac{Q_{REPEX}}{Q_{CONN}} \qquad (4)$$

which makes clear that the relative values of the elasticities are based on the underlying marginal costs of the components of the index, as well as the relative volumes of activity for each of the components.

Turning to Ofgem's current approach, it is important to note that the use of industry cost shares as weights, combined with the inclusion of a single CSV measure in the regression model, places an implicit restriction on the relative elasticities of the individual components of the index. The current approach can be written as follows:

$$Ln\ Totex = a_1 + b_1\ Ln\ CSV + error \quad (5)$$

which can be re-written as:

$$Ln\ Totex = a_1 + b_1\ Ln\ (CONN^{0.37}.REPEX^{0.63}) + error \quad (6)$$

and in turn as:

$$Ln\ Totex = a_1 + b_1.0.37.Ln\ CONN + b_1.0.63.Ln\ REPEX + error \quad (7)$$

From equation (7) it is clear that a model comprising a single CSV measure can actually be re-written to look like one with the individual components included separately. However, the difference between equation (7) and equation (2) is that equation (7) is restricted. The ratio of the elasticities on CONN and REPEX is forced to be equal to the relative cost share weights: that is, 0.63 / 0.37 in this case.

Thus, whereas the unrestricted model sees the relative elasticities being based on the underlying marginal cost of each of the activities, and the volumes of those activities (Equation (4)), now the relative elasticities are based on the relative weights, which are in turn based on the cost shares:

$$\frac{b_1.0.63}{b_1.0.37} = \frac{0.63}{0.37} = \frac{UC_{REPEX}}{UC_{CONN}} . \frac{Q_{REPE}}{Q_{CONN}} \quad (8)$$

where UC represents the unit (or average) costs of the different activities.

Equation (8) looks similar to Equation (4), with the difference that the relative elasticities are based on the unit costs (average costs) of the different activities, rather than the marginal costs.

In theory, the cost relationship / elasticity should be based on the marginal cost of changes in outputs and not the unit (average) cost – unless the two are equal of course. Therefore the use of a single CSV measure in the Totex cost model is overly restrictive because the relative elasticities are based on the relative unit (average)

costs of the two outputs, rather than the relative marginal costs, which may well be different (or at least, the model should permit the possibility that they are different)[1].

The restriction implied by Equation (7) can of course be tested statistically. If this restriction is rejected, it could imply that a less restrictive model ought to be considered. In other words, such an outcome might suggest that the industry cost share weights used are not appropriate and that an unrestricted model, including all the components of the CSV separately, would let the data speak about the relationship between costs and the components (thus revealing effectively a different set of weights). That said, including all of the components of the CSV directly in the model in an unrestricted way might lead to problems of multi-collinearity, and produce elasticities on the individual components that are considered implausible or hard to interpret. It is important not to overplay this latter problem, however, as it is a standard problem / trade-off in all cost modelling work.

## 3. An example

Table 1 shows an unrestricted model based on the actual CSV components used for the Totex model (based on 6 years of data for the years 2013/14 to 2018/19, i.e. RIIO-GD1 actuals). Table 2 shows the weights based on industry expenditure used to construct the Totex CSV measure for comparison.

Statistical testing indicates that the restricted model is not supported by the data. The unrestricted model (Table 1) allows the data to reveal the influence that the components have on costs in an unrestricted way. This model is then compared to a model that imposes the restrictions that the elasticities on the components reflect the industry expenditure share weights used to compute the CSV measure (in terms of the simplified equations presented above, this is the equivalent of comparing the restricted equation (7) with the unrestricted equation (2)).

[1] This discussion has focused on the Cobb-Douglas functional form, which is the one that is most commonly used in economic regulation. In the more complex translog case again the use of a CSV imposes constraints on the underlying coefficients of the model.

**Table 1: Ln Totex vs Components of CSV**

| Variable | Coefficient |
|---|---|
| Constant | -3.399*** |
| LNMEAV | -0.368 |
| LNEMER | 0.612** |
| LNEXT | 0.0864 |
| LNMMEAV | 0.245* |
| LNREP | 0.159** |
| LNCONNW | 0.061 |
| LNMAINW | 0.002 |

R-bar squared = 0.97395. Number of observations = 40. *** = statistically significant at the 1% level. ** = statistically significant at the 5% level. Notes: LNMEAV = Ln MEAV; LNEMER=Ln Emergency CSV; LNEXT = Ln External condition reports; LNMMEAV = Ln Maintenance MEAV; LNREP = Ln Repex workload; LNCONNW = Ln Connections workload; LNMAINW = Ln Mains workload. Year dummies were also included (not shown)

**Table 2: Actual Totex CSV weights**

| Variable | Weight |
|---|---|
| MEAV | 0.369 |
| Emergency CSV | 0.055 |
| External condition reports | 0.060 |
| Maintenance MEAV | 0.060 |
| Repex workload | 0.416 |
| Connections workload | 0.023 |
| Mains workload | 0.016 |

Since statistical testing convincingly rejects the restriction, this casts doubt on the approach of using a single CSV measure in the econometric model (with weights based on industry expenditure data). However, the unrestricted model in Table 1 shows that only some of the components of CSV are found to be statistically significant, and in one case a negative elasticity is reported. The regression results therefore imply quite different relative weights to those used within the CSV, but at the same time there is considerable uncertainty over the results given the lack of statistical significance and the presence of negative signs. Thus, there are statistical reasons for doubting the use of a single CSV measure, but at the same time regressing Totex on all of the components produces results that are hard to interpret without further work.

A similar finding emerged in previous work carried out by Ofgem and its advisors as part of GD1[2], thus motivating the continued use of the single, CSV regression approach; though for ED1[3] the econometric approach was also used.

## 4. Concluding remarks

It is worth recalling that the objective here is to explore the use of econometric techniques that let the data reveal the underlying cost impacts of the different activities included within the CSV composite variables. In principle this approach can provide important, new information on cost-output relationships in the industry. The use of a CSV measure, based on industry expenditure shares, is restrictive as explained above, and the econometric approach in principle allows us to relax (and test) those restrictions. The key challenge in this case is whether there is enough data to operationalise the approach.

To some extent the findings and discussion in this note raise issues that are standard in any model selection decision, and are repeatedly played out in regulatory cost assessment studies. It is often the case in econometric cost modelling work that there may be several variables that are expected to affect costs, but their simultaneous inclusion in the model may produce estimates on some of the variables that are statistically insignificant and / or implausible – often because of relatively small sample sizes. Thus, there can be a motivation to deliver a more parsimonious model by dropping some of the variables from the model. This can even be the case when statistical testing rejects the simpler model (for example, in choosing a Cobb-Douglas form over a translog). In this respect the current problem is no different.

At the same time, there can be an argument for leaving all of the elements in the model, in order to ensure that the estimates of inefficiency (within the residual) are purged of the effects of all of the variables. Under this approach, it has to be accepted that not all the parameters will individually make sense, but that we capture the overall impacts of the variables. In many railway applications, for example, it is accepted that the elasticity of costs with respect to freight traffic may be small and insignificant, given that it is often the case that freight makes up a small share of traffic. Thus it is not necessarily the case that a model should be rejected just because some of the parameters are not statistically significant. In the present case some degree of collinearity between the variables would be expected, particularly since changes in the MEAV variables will partly reflect changes in some of other measures. It is usually the case, however, that there is no definitive answer to which

[2] https://www.ofgem.gov.uk/ofgem-publications/48198/gd1initialproposalsstepbystepguidefor-cost-efficiencypdf

[3] https://www.ofgem.gov.uk/sites/default/files/docs/2014/11/riio-ed1_final_determination_expenditure_assessment_0.pdf (Appendix 5)

model will be most appropriate, and regulators would need to exercise judgement, taking account of a range of model selection criteria.

The preliminary work carried out seems to imply that there is a difficult trade-off between continuing with the current approach (which seems to be rejected based on statistical testing), and adopting a more flexible approach that produces results that at first sight may be hard to interpret (in terms of the elasticities), but which fits the data better. This is a standard trade-off in econometric work and is hard to resolve; it ultimately requires some judgement to be applied.

Given that similar issues have emerged in previous work by Ofgem and its advisers, and, given the relatively small sample size, further work may not yield useful results. However, I do consider that it would be useful to investigate different model specifications containing alternative combinations of variables[4], trying different time periods of data, and also comparing the efficiency scores across different models for completeness. It may be that by including only a sub-set of variables in the model, a more plausible set of elasticities might emerge, with increased statistical significance and the avoidance of negative signs. Indeed, given that some measures (such as MEAV) overlap with other output variables, there may in any case be a justification for dropping some of the variables. Alternatively, if it is desired to leave all variables in the model together, there could be an argument for accepting such a model, and deriving efficiency estimates from it, even if some parameter estimates are small and insignificant, or even counter-intuitive. Such an approach would though require careful investigation of the implications for individual companies, and reassurance that the results can reasonably be explained, for example by correlations in the variables.

It would also be interesting to compute the marginal costs for different activities based on the elasticities reported in the unrestricted model, and check these for plausibility. This could encourage a discussion on the level of marginal and average costs for different activities in the industry. It might also be considered whether “sub-company” data could be used to expand the sample size and thus improve understanding of the underlying cost relationships.

Whether or not the end result is to retain the existing method, new insights may emerge from the analysis, and at least the alternatives will have been fully tested and documented; and the reasons for the selection of one approach over the other set out clearly. If it is desired to retain a model with a single CSV variable in the model, the more detailed regression work could potentially be used to refine the weights used in computing the CSV, or at least to prompt further debate and understanding about cost relationships in the industry.

[4] Potentially alternative functional forms could be considered.