November 7, 2022

Dr Jonathan Bourne
Little Bear Consulting
London
Phone: +44 07958 717 256
Email: jonathan.s.bourne@gmail.com
Linkdein: Jonathan Bourne

Arne Hanto Moen
Øvre Eiker Energi
Skotselvveien 29
3302 Hokksund

Dear Arne Hanto Moen,

This is a short discussion into the "Forslag om endringer i modell for kostnadsnorm. Korrigering for rammevilkår", it aims provide a review into the report in much the same way as an academic peer review, a process the authors are familiar with. The significance of changes to the 'modell for kostnadsnorm' are that the amount of money a power distribution company can receive will change (inntektsreguleringen). As such it is important that any proposed model is accurate, reliable, and grounded in appropriate theory.

Given the information and explanation presented in the report I do not think that the model in "Forslag om endringer i modell for kostnadsnorm. Korrigering for rammevilkår" reaches the required standard in the three previous areas and should not be used.

In addition the report is poorly structured, and lacks important methodological details. Whilst I appreciate that the target of the work is not aimed at statisticians, it still needs to be able to be evaluated. Indeed given that non-statisticians are going to attempt to evaluate it extra care needs to be given to provide clear and straight forward explanations of the methods and maths used and their significance.

1 Discussion

The points are broken up into three sub-sections. Two subsections provide an analysis into core areas of the model and its interpretation, the third section provides comments on specific points in the report.

1.1 understanding the model performance

The predictive power of the model is low. The model has R^2 value of 0.42 meaning that 58% of the variance is unexplained. Figure 1 provides a visual example of what this level of noise looks like, using the assumption that the errors are normally distributed and homoskedastic. From the report it is not clear whether this noise is a function of management or unknown variables not yet included in the model, or both. This is important as if the deviance from the line is purely management quality this will need to be clearly justified as it makes strong statements about how each company is run. If the variance is from some unknown variables, then basing income off such a model runs a very real risk of over or under remunerating an organisation by a large amount.

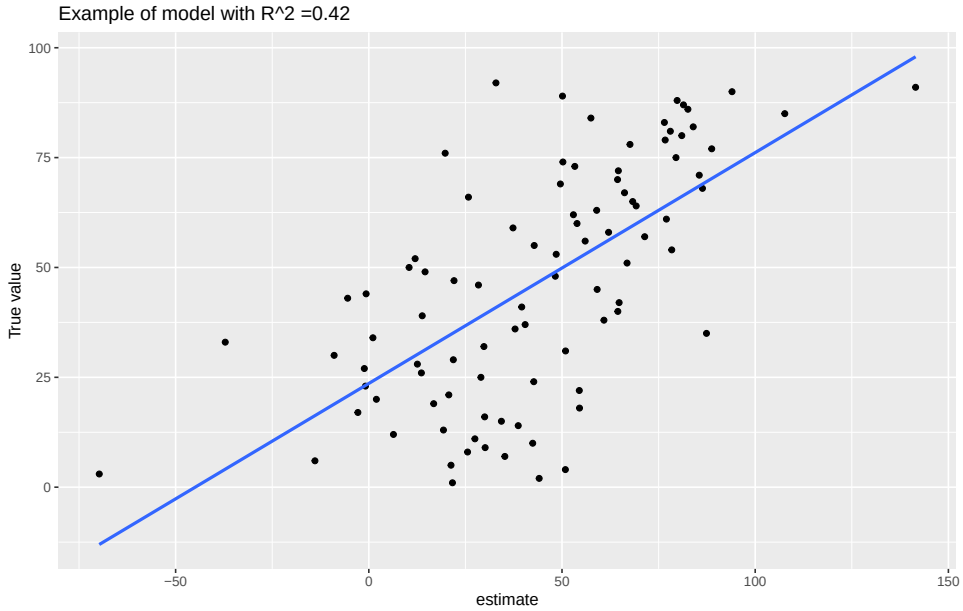

Figure 1: The figure gives a visual example of the accuracy of the proposed model. Is the deviation from the line missing variables? Mangement effectivness? both?

Although the authors discuss the stability of the model coefficients using a time series. They do not seem to measure it using cross-validation. Cross-validation is a standard measure of understanding coefficient stability and model fit. To get an understanding of cross-validation consider the example of 10-fold cross validation. For 10-fold cross validation the data is randomly split into 10 groups labelled 1-10. In fold 1 a model is trained using groups 2-10 and tested on group 1. Fold 2 trains a model on all groups except group 2 which is used for testing. The process continues until all groups have been used for testing, this creates ten models with ten sets of coefficients and ten sets of performance metrics.

Any model is a proxy for some unknowable or latent truth. The coefficients are approximations of the 'true' coefficients and the model performance an estimate of the true performance. Cross-validation gives an indication of the range within which these true values lie. If the distribution of the coefficient or performance values is small it means that the model is not over fit or overly influenced by a small number of points.

To test the stability of a model of the quality created by the authors I used a process called repeat 10-fold cross validation. This is simply the 10-fold cross-validation repeated multiple times on different random

Figure 2: The histograms show the distribution of model performance/error. It can be seen that given the range of data is between -25 and 150 the models are unreliable.

splits. I did 100 random repeats of 10-fold cross validation. This created 1000 unique models from the same data.

The results of the 1000 models are shown in Figure 2, the average root mean square error (rmse) is 26 whilst the mean R^2 is 0.54 [1]. The 95% confidence interval, that is the values for which 95% of the data falls between is between 16 and 37 for rmse and between 0.11 and 0.86 for R^2. Such a large range of possible performance results indicates that any model produced by these data is likely over-fit, being influenced by a few points.

If we assume that the deviance from the model is due to missing variables that explain the performance and not simply effective management, the model proposed by the author will result in miss-allocation of funds worth millions if not tens of millions of kroner.

1.2 Understanding model performance relative to the weighting

An interesting aspect of the application of the results of this model are that it comes in two parts. The performance of the model itself, and the weighting of the model overall. The model performance is a measure of how well the data fits the dependent variable and is a strictly statistical measure. The weighting of the model is how well the dependent variable matches some latent or observable operational costs, from the report this appears to be decided using non-statistical methods.

The separation of the model performance and the weighting implies that the true cost of operating a distribution network is not known/unknowable and thus an observable proxy is required. The weighting which effectively evaluates the fit between the observed and latent values appears to be decided by expert analysis. There is nothing inherently wrong in using non-statistical methods to weight the dependent variable against the latent operational costs. However, it should be made clear what is going on. There also needs to be an explanation of why the model dependent variable is not the same as the latent variable and the basis for the weightings given. This latent truth can be related to the multiple model approach. None of this is present in the report. Figure 3 provides a diagram of what this implied latent structure looks like.

Figure 4, shows the position of the proposed model on the R^2 and latent weighting axis. The model

[1] It should be noted that the average R^2 for this model is substantially better than the 0.42 recorded in the report

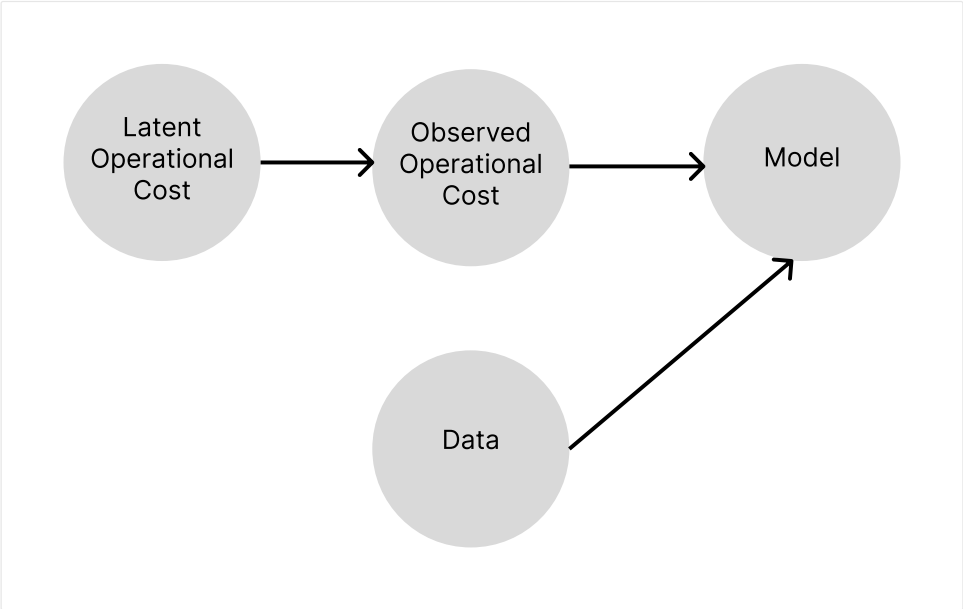

Figure 3: The weighting of the model implies latent truth for which the dependent variable is only a proxy

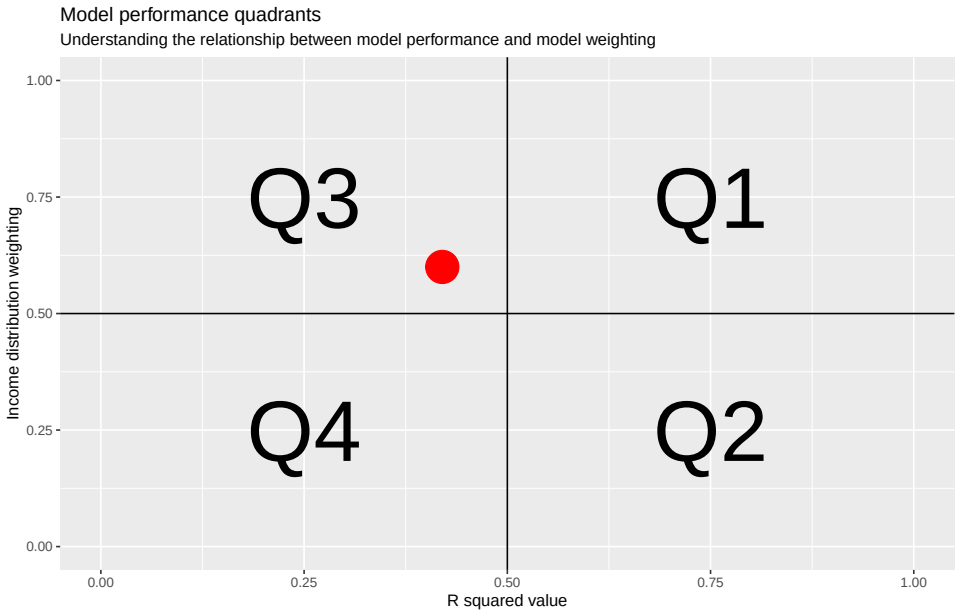

Figure 4: The 4 performance quadrants of the model, combining both the expert weighting and the statistical model. The red point is the current model

weighting is 0.6 whilst the R^2 is 0.42. The figure is broken into four quartiles with Q1 being the best and Q4 being the worst. The four quartiles can be qualitatively evaluated to give them meaning.

- Q1: The data explains a large proportion of the variance of the dependent variable. In addition the dependent variable is a good match for the operating cost of a distribution network
- Q2: The data explains a large proportion of the variance of the dependent variable. However, the dependent variable is a poor match for the latent operational costs.
- Q3: The data is a poor fit and much variance is unexplained. However, the dependent variable is a good match for the latent operational costs
- Q4: The data is a poor fit for the dependent variable and the dependent variable is not a good match for the laten operating costs.

In an ideal world we would be able to accurately predict the latent costs using a model with a high R^2. The current model sits in Q3. That is the model is not very good, but the dependent variable is considered a good fit for the latent costs.

This then raises further questions. What does the costs used to create the model represent if not the operation costs? I am lead to believe that the costs used in the mode include IT and administrative costs. If these are included it would explain why the output variable is not considered a perfect match for the cost of operating the physical infrastructure. However, why are costs unrelated with the operation of the physical infrastructure included? Does this explain the poor fit for the model? Is the model actually supposed to include some value of management efficiency?

Ultimately the question of what is being measured and how well it represents the true latent thing to be measured is probably more consequential than the inclusion of many of the variables. However, like much else this is not discussed.

1.3 General comments

- In the Forward to the report the authors state that the kostnadsnorm modellen accounts for 60% of the income framework and it is proposed to increase this to 70%. This indicates that there is a strong belief that the model captures a substantial amount of the cost variation of distribution companies and thus the deviation from the model is primarily due to management differences. Considering the poor performance of the model this is a strong position to take. The authors show that the model explains just 42% of the total variance in cost. To increase the importance of such a model, seems beyond explanation, and should be clearly justified.
- The final paragraph of page nine begins, “Vi er enige i at statistisk sammenheng ikke bør være det eneste kriteriet for hvilke variabler vi inkluderer i modellen.” This is a remarkable statement and risks bringing the validity of the entire model into question. Including variables in a model with no principled framework does not add value to a model only noise, and with the small quantity of observations in the dataset there is a real risk of generating spurious results. If the authors want to include expert opinion then they should use an appropriate framework such as Bayesian statistics, this will allow them to adjust the priors of the model in a way that is clear and interpretable.
- Too much of the report is concerned with the views of various distribution companies. These views are important but not central to the paper. The authors should move such views to an appendix and instead state the assumptions they are using for the model using clear and logical argument. Mixing in a large number of views makes understanding the fundamental assumptions difficult.
- The method does not contain enough actual method. Instead it is a mixture of more discussion on the views of various companies, method and results. Given the significance of this model it is critical that the method is clearly explained and utterly distinct from the results. As an example, section 4.3 discusses the variation of the coefficients across time, however the model is only defined in section 5. This makes no sense and needs to be changed.

- Section 4.2 describes combinations of variables. It is not clear from the description whether this refers to the interaction between variables or the correlation between variables. PCA does not 'combine' variables rather it finds a rotation in the parameter space which makes the resultant variables uncorrelated. This is an important distinction as the resultant variables may still have valuable interactions which need to be accounted for. The model described in the report does not have interaction terms.It should also be noted that using PCA as a variable reduction technique results in information loss, this is not explained in the work, no elbow plot or measure of information loss is described.

- Regarding PCA the authors correctly identify the risks associated with creating a linear model with highly correlated variables, they use these risks to justify the use of PCA. This is entirely reasonable. But why then are there three separate models built using the same output variable, the models of "Trinn 1, Trinn 2, and Trinn 3? if the final value is a linear combination of the three models then the authors have effectively created a single linear model and risk inducing multi-co-linearity anyway. Such a decision brings into question the strategy of having three separate models that 'correct' for each other, which itself is a strange decision to take, and for which I have not seen a clear explanation (although it may exist).

- One must be careful when adding in new variables for such a small dataset, that the variables have meaning across a reasonable proportion of the observations. For example, if a binary variable is only positive for 3 out of 100 observations, it should not be included as there is a strong risk of inducing bias. This is a reason for performing proper cross-validation (see section 1.1)

- It would seem intuitive that the biggest drivers of operating a power grid come from the topological and structural properties of the network, however the model explicitly rejects such variables on the grounds that companies have a degree of control over this. No explanation beyond that the authors are in agreement with "selskapene ikke bør kunne påvirke rammvilkårsvariablene" is given. That a company has influence over such structural properties, is this because these are influenced by the quality of management? Or because it may encourage gaming the system? This is not explained at all. It seems highly unlikely that a company would engage in substantial infrastructure projects to game a system such as this. Beyond the fact that the algorithm can be changed, the risks of such infrastructure projects are high and the rewards, stemming from the model, marginal. To argue that a cost model for a transmission company should exclude network features because the power company can influence it, is like suggesting that a cost model for a logistics company should not include vehicles and kilometres travelled.

- There is not comparison or discussion of the current model relative to the previous ones. What is the statistical or predictive basis for using this model over previous ones?

- The multistage model is confusing to understand. the DEA assumes all deviance is due to inefficiency but then a parametric model is added in as the DEA is known to have errors, the third model is added in on top for other reasons. This makes it very hard to get a full understanding of what is happening and how well it may be representing the true inefficiencies. Overall it feels like this report is not a self-contained piece of work but requires the reader to go through other documents by the authors in order to understand what is happening.

- I don't know how common it is to mix DEA with linear regression, but it does make interpretation more difficult. Although COLS has it's drawbacks it is interpretable and can be adjusted much more easily. Although I am not familiar with DEA it seems to me to require a very rigorous theoretical framework for including variables as the quality of fit cannot be measured, a requirement that does not apply to COLS. It would be interesting to understand why the authors choose to use this approach over COLS considering the steps that need to be taken afterword.

- It would also be interesting seeing the performance of an OLS model for Trinn 1 and how well it performs. It would also be interesting to see these variables and the variables from Trinn 2.

2 Conclusion

This review has highlighted a range of issues which affect 'Forslag om endringer i modell for kostnadsnorm. Korrigering for rammevilkår'. I am not saying there is anything fundamentally wrong with the approach, however, there are substantial issues that need to be resolved before such a model can be applied in a business context.

Finally, it appears a process of reviewing Trinn 1 has begun, I think that it may be an opportunity to pause any further changes to Trinn 2 and review the entire process. The review could be performed with support from those with the relevent expertise from UiO, NTNU or other appropriate institution.

Sincerely yours,

Dr Jonathan Bourne