A Brief Note on the CMA's Implementation of LASSO.

4th November 2025.

Dear CMA Panel,

I am writing this brief note in response to the use and implementation of LASSO - and more generally the provisional findings – relating to the appeals made by water companies concerning Ofwat's PR24 Final Determinations (FD) for the 2025 to 2030 price controls.

I am a doctoral researcher at Loughborough University and awaiting my viva voce examination having submitted my PhD thesis in September. My research focused on developing robust cost assessment methodologies for performance measurement of wastewater systems and was more generally interested in determining the cost of wastewater service provision. Moreover, I have studied and implemented machine learning estimators including LASSO (and its many variants) over course of my PhD research. I fully appreciate the potential benefits of employing such techniques, while I have also considered in great detail the potential downsides through extensive simulation exercises and real-world data applications. Therefore, I offer this submission to highlight some concerns with regards the implementation of these models, with the belief that I may have some valuable insights.

In the interests of transparency, I would like to note that Anglian Water Services partially funded my PhD research – hence my interest in these proceedings and understanding of the regulatory process. However, I am not currently in receipt, nor do I expect to receive, any funding or payment from Anglian Water based on this submission – I am engaging with this call for views on the CMA's provisional determinations solely now as an interested onlooker, albeit with a reasonable level of expertise on many of the related issues.

As an additional preliminary note, I believe that the CMA should be commended for showing ingenuity in exploring econometric or machine learning methods beyond the traditionally relied-upon approaches. However, it is also my belief that the use and implementation of LASSO in this specific context is inappropriate – in fact, I do believe that such an exercise can be very valuable for a regulator such as Ofwat as an initial first step of the regulatory cost setting mechanism before the relatively arbitrary restrictions are set on the fully available (master) datasets. Such a circumstance (a larger dataset) is precisely the time where LASSO and other machine learning techniques exhibit their virtue. However, for a number of reasons to be discussed below, the use and implementation of LASSO is inappropriate within the current very limited problem bounds presented to the CMA.

Section 1 – LASSO

Conceptually, LASSO is an econometric tool which serves as a method of dimension-reduction. That is to say that LASSO is most suited to situations where large datasets (containing many possible variables, and a large number of observations) are available and the user is interested in identifying the best subset of the available data in order to predict - with reasonable accuracy – an outcome variable of interest, e.g., BOTEX. This best subset of the data is often referred to as the 'optimal' selection of variables, or the subset of variables containing the most explanatory power.

Notwithstanding the desirability of large datasets, as is often the case in regulatory settings, only much smaller datasets are typically available in practice[1]. In such cases, the use of LASSO should therefore be approached with caution. The following criteria are helpful:

Criterion 1 - Thoughtful consideration in identifying an initial set of candidate variables.

Criterion 2 - Caution in accepting LASSO's 'optimal' selection of variables.

Given that LASSO is a data-driven technique, it is prudent in a regulatory setting to use an informed initial list of candidate variables from which we allow LASSO to 'choose' from. This initial informed list would in practice include many potentially powerful cost drivers – such as those reasonably aligned with engineering rationale, economic theory, customer dynamics, physical geography, network characteristics, and company operating environments. Just as importantly, it is then prudent to assess and scrutinise the optimal selection of variables that LASSO identifies. Ensuring that this optimal selection - from the perspective of a data-driven metric - aligns with engineering rationale, economic theory, and industry experts expectations (etc.) is just as important in a regulatory setting. As an illustrative example, when predicting costs of wastewater services LASSO could in theory propose an 'optimal' selection of variables that only includes the various types of sewers employed (surface water-only, foul-only, rising mains, combined sewers, etc.) – however, this hypothetical optimal selection says nothing about the volume of output, treatment complexity, service area, population density, etc. As such, these prudent criteria, when used together, ensure that the LASSO tool is not simply used as an under-informed fishing exercise – the kind of fishing exercise undertaken by marketing and sales companies who use Big Data to identify potential customers.

Alas, a problem arises with the CMA's implementation of LASSO. Firstly, the list of candidate variables from which LASSO is required to select from is overly restricted. The CMA limits this list of candidate variables to include only the variables used by Ofwat in the PR24 FD models and three new

[1] In practice, Ofwat's PR24 Cost Assessment Master Datasets for Wholesale Water and Wastewater Base Costs might be considered 'wide' as they contain hundreds of variables, yet they are also relatively 'short' as there are few companies observed in any given year of data.

(additional) variables in the case of treated water distribution (TWD) and water resources plus (WRP) models, while only two of these additional variables are considered in the case of the wastewater (WW) model.

Given the CMA's function in the regulatory cost-setting process, it appears as if the CMA is really using LASSO as an attempt to inform dimension-increase of Ofwat's PR24 FD cost models. In other words, LASSO is being used to identify which of the newly considered variables can be added to (or perhaps at worst substituted into) Ofwat's list of chosen variables; rather than using LASSO as a tool for dimension-reduction of a bigger (but nonetheless, 'informed') less-restricted dataset containing many potentially powerful explanatory variables – the type of dataset described by **Criterion 1**. This might seem like a subtle difference, but in my experience runs counter to best practice when using LASSO or other machine learning variable selection techniques. This critique would be abundantly evident in a scenario where LASSO returned a selection of variables that decimated Ofwat's FD models, acceptance of such a selection would bring the PR24 cost models - and the cost allowances assigned to non-disputing companies - into disrepute.

Section 2 – Candidate Variables and the Optimal LASSO Selection

The proposed selection of variables under LASSO detailed in Table 4.1 of paragraph 4.58 on p.54 of the Provisional Determination (specifically, the column titled "*Cost drivers included*") raises some interesting questions and even sets a murky precedent if accepted in the immediate term. Given time constraints, I have selected the following as examples:

- Firstly, efforts to include input prices (proxied by wage and energy indices) should be commended as economic theory and rationale both point to these as legitimate cost drivers. Nonetheless, two aspects of the CMA's implementation are difficult to comprehend[2].
    - Ignoring natural log transformation notation for now, the multiplicative terms constructed for energy and wage index with a chosen scale driver in TWD and WRP models (as input index * length of mains) and for the WW model (as energy index * pumping capacity, and wage * load) are poorly justified and frankly, very difficult to justify. Given that scale is already accounted for within each of the models, this

[2] An additional (third) issue relates to the uncertainty around whether natural log transformations have been imposed on wage and energy indices or not - Table 4.1 (paragraph 4.58 on p.54 of the Provisional Determination) suggests no such transformation takes place and Table D.2 (paragraph D.14 on p.24 of the Provisional Determination Appendix) corroborates this, while Table D.3 (paragraph D.14 on p.24) and Table D.4 (paragraph D.18 on p.25 of the Appendix) both suggest that natural log transformations have in fact been used. If natural log transformations are only used selectively this should be justified as this is also likely to introduce bias to the LASSO estimator. This issue is entered as a footnote to give some benefit of the doubt but nonetheless requires clarification.

construction of multiplicative terms to capture higher input usage for bigger companies is completely unnecessary, and in fact obscures the interpretation of the scale driver coefficients.

- Notwithstanding the point made above, it is apparent that the energy index * scale driver parameters in TWD, WRP, and WW models are identified as important for inclusion by LASSO. However, the wage index * scale driver is only identified and included in the TWD model. This effectively suggests that wages are not a legitimate cost driver in WRP and WW functions. The selective exclusion of controls for wages in the WRP and WW functions is again very difficult to justify and does not appear to be addressed in any great detail – this is a representation of the **Criterion 2** listed in Section 1, particularly when using LASSO in an applied regulatory setting: the final selection by LASSO should be scrutinised appropriately.

- The inclusion of multiple density measures in the candidate variable list for each model (the summation of the "*Cost drivers included*" and "*Cost drivers dropped*" columns of Table 4.1) is both inconsistent with Ofwat's modelling principles and past precedent, e.g., strict adherence to a single scale variable and the strict avoidance of multicollinearity. While the expected signs and magnitude of density variable coefficients are understandably difficult to determine a priori, the estimated results and subsequent interpretation of density variable coefficients given by the CMA in Appendix D (paragraph D.17, on p.25 of the Provisional Determination Appendix), highlights an major incoherency with the principles applied by Ofwat in designing the suite of models used for PR24 FDs. In Table D.4 (paragraph D.18, on p.25 of the Appendix), density variables are seen to exhibit coefficients with opposite signs and not too dissimilar magnitudes – the explanation of such is that this contradictory effect is "consistent with the view that the impact of density on costs is unclear". Frankly, this interpretation has the potential to set a very murky precedent; if this justification were rolled out every time an econometrician saw an instance of model misspecification, it could severely undermine the consistency and credibility of model-based cost assessments. Moreover, one of the primary robustness checks used by Ofwat in PR24 is whether estimated coefficients are of the right sign and of plausible magnitude. Therefore, given this incoherency, I would suggest that either the model development undertaken by Ofwat was unnecessarily restricted by this robustness check, or the present inclusion of multiple density measures requires reconsideration as it violates the same robustness check.

- The candidate variables for the TWD model (the summation of the "*Cost drivers included*" and "*Cost drivers dropped*" columns of Table 4.1) indicate what is at best a conceptually inconsistent approach, and at worst an *ad hoc* and contradictory approach to model specification and identification of candidate variables.
    - Once again ignoring log transformation notation due to the uncertainty raised in Footnote 2; in the TWD candidate variables 'length of mains' enters as a scale driver (the justification for which might be as simple as: more pipeline equals higher cost) – implicitly, this says that costs increase proportionally with mains length. Separately, a ratio term of the number of properties per length of mains enters the candidate variable list – implicitly, this says that we are not so much concerned with the number of properties or the precise length of mains, but rather we are only concerned with the ratio of properties to length of mains. Then, based on the CACs, a third type of variable enters the list of candidates as: the length of mains multiplied by the energy (or wage) index – implicitly, this then says we are interested in both the individual levels of mains length and energy (or wage) indices, and the relationship between the components. Ordinarily, the use of simple indicators, ratios, and multiplicative terms would not necessarily be an issue, however, the scale driver in this case (length of mains) is subject to three different sets of assumptions, and these assumptions contradict one another. Moreover, it can be said that this issue only exists because of the unnecessary construction of multiplicative terms between input indices and scale drivers (length of mains in this case). In turn this issue raises complications with regards introducing unnecessary multicollinearity, identification issues, and interpretation of coefficients – all of which should be basic robustness checks.

To conclude, I would strongly encourage the CMA to reconsider both its use of LASSO within the very limited problem bounds presented and failing that – at a minimum - its implementation of LASSO. The issues highlighted herein cast into doubt the likelihood of producing unbiased estimates of efficient costs using the proposed models.

While every effort has been made to ensure accuracy, any errors or omissions in this submission are entirely my own.

Yours faithfully,

Fionn Cliffe