ISSN 1973-2201 STATISTICA, anno LXXXIV, n. 2, 2024
DOI: https://doi.org/10.60923/issn.1973-2201/20656

DISCUSSION OF THE PAPER "CONNECTING MODEL-BASED AND MODEL-FREE APPROACHES TO LINEAR LEAST SQUARES REGRESSION" BY LUTZ DÜMBGEN AND LAURIE DAVIES (2024)

Efthymios Costa [1]
Department of Mathematics, Imperial College London, London, United Kingdom
Ioanna Papatsouma
Department of Mathematics, Imperial College London, London, United Kingdom

The paper by Dümbgen and Davies (2024) offers a fascinating perspective on p-values and hypothesis testing within the context of linear regression. The main idea of the paper is effectively communicated to the reader, presenting a model-free interpretation of p-values, a challenging and often contentious topic, especially for early-career statisticians. The presented approach makes the paper an engaging read.

One of the key contributions that stands out is the focus on model-free interpretations. A plethora of contemporary statistical methods are reliant upon model assumptions which, when unmet, may produce inconsistent results. Therefore, model-free approaches can be attractive for assessing how good of an approximation a model is for a given set of data. This is particularly valuable in the case of high-dimensional regression. In this specific example, if the set of covariates is given by $\mathscr{S}$ and the 'assumed true' (by the words of Tukey, 1993) unknown subset of predictors that have generated the data is $\mathscr{S}_0$, the task is to get as close to $\mathscr{S}_0$ as possible. Dümbgen and Davies (2024) show how their approach yields subsets of covariates with lower sums of squared residuals than the established and well-known lasso method.

It is worth pointing out that several advances to the lasso methodology have been proposed. Selective inference methods have emerged to mitigate selection bias in high-dimensional settings (Benjamini, 2020). Furthermore, recent work by Zrnic and Fithian (2024) introduced 'locally simultaneous inference'. Their approach consists of focusing only on subsets of variables or models that could plausibly, rather than potentially be asked. However, they acknowledge that an understanding of the selection mechanism for $\mathscr{S}_0$ is crucial. It would be of interest to find out how these selective inference tech-

[1] Corresponding Author. E-mail: efthymios.costa17@imperial.ac.uk

niques, designed specifically for high-dimensional regression, compare to the model-free approach of Dümbgen and Davies (2024).

Finally, we thank the authors for engaging with the deeper philosophical concepts of 'statistical truth' and 'ontology'. These reflections are especially relevant today, as big data continues to reshape the statistical landscape (Dunson, 2018). We look forward to seeing future developments stemming from this thought-provoking work.

References

Y. Benjamini (2020). *Selective inference: the silent killer of replicability*. Harvard Data Science Review, 2, no. 4.

L. Dümbgen, L. Davies (2024). *Connecting model-based and model-free approaches to linear least squares regression*. Statistica, 84, no. 2, pp. 65–81.

D. B. Dunson (2018). *Statistics in the big data era: failures of the machine*. Statistics & Probability Letters, 136, pp. 4–9.

J. W. Tukey (1993). *Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper*. Princeton University, Princeton.

T. Zrnic, W. Fithian (2024). *Locally simultaneous inference*. The Annals of Statistics, 52, no. 3, pp. 1227–1253.