ISSN 1973-2201 STATISTICA, anno LXXXIV, n. 2, 2024
DOI: https://doi.org/10.60923/issn.1973-2201/22200

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

Christian Hennig [1]
Department of Statistical Sciences "Paolo Fortunati", University of Bologna, ALMA MATER STUDIORUM, Italy

1. Assumed models do not need to be true

A major feature of Dümbgen's and Davies's (DD) paper (Dümbgen and Davies, 2024) is the provision of probabilistic theory for least squares regression that does not rely on any model assumptions concerning a data generating process (DGP) behind the data. *P*-values of the classical *F* test, based on model assumptions including linearity and Gaussian errors, are given a model-free meaning.

Although the discussion paper itself does not criticise the model-based approach very explicitly, Laurie Davies does so in closely related work (Davies, 2024), where he argues that it makes little sense for statisticians to behave as if a certain probability model were true given that they know perfectly well that "all models are wrong" as George Box stated.

In Hennig (2023) I advocate an attitude to statistical models that acknowledges explicitly that such models are potentially helpful thought constructs, and that their job is not to be "true". Different from DD, however, my focus is on a re-interpretation of the classical results involving assumptions regarding the DGPs that brought forth the observations, whereas DD's model-free view uses a probability distribution to create artificial random variation that is separated from the observed data.

Having classical theory based on model assumptions regarding methods of statistical inference does not mean that we need model assumptions to be "true" in reality. It rather means that we can investigate the workings of the methods in an artificial benchmark situation in which we can control the mathematical truth that a method is meant to get at. Even though the model assumptions will not hold in reality, it is reasonable to use

[1] Corresponding Author. E-mail: christian.hennig@unibo.it

methods that are guaranteed to perform well in such an idealised situation. Such theory has also been very stimulating for the creation of statistical methodology, often through optimisation.

Of course a theoretical performance guarantee of a method under certain model assumptions does not guarantee a good performance where the model assumptions do not hold, and it is rather subtle and strongly dependent on the situation whether such a method will perform well or not. But in any case the model-based theory contributes to the understanding of the method's characteristics in a valuable way.

Model-based simulations as used in Sec. 3.3 of DD's paper are informative in the same way; even though DD derive theory that does not come with model assumptions for the DGP, it is of interest to ask how well a method performs in a situation where we know what it should optimally do, and statistical models just provide such situations. Without assuming "true" parameter values, for example there is no such thing as a type I or type II error or mean squared error, but these concepts help to measure how well our methods do.

2. Interpretation of equivalence regions and confidence sets

It is a well known issue with confidence sets that the confidence level $\beta = 1-\alpha$ is a performance characteristic referring to repeating the real experiment of interest infinitely often, which in reality of course cannot be done.

After observing data $\mathbf{x}$, say, even assuming that a true parameter θ exists, the probability that θ is in a confidence set $C_\beta(\mathbf{x})$ is not β, although practitioners tend to interpret it like that.

DD do not define equivalence regions in a fully general manner, they rather present examples. In general I guess that their interpretation should be something like this ($C_\beta(\mathbf{x})$ denotes the equivalence region here):

Let $\{P_\theta, \theta \in \Theta\}$ be a set of parametric distributions with parameter set Θ. W.l.o.g. let the statistic S be a non-negative function of $\mathbf{x}$ and θ so that smaller values of S indicate a better "fit" (in some sense to be defined) of $\mathbf{x}$ by P_θ. Let $C_{S,\theta,\beta} = [0, q_{\theta,\beta}]$ with β-quantile $q_{\theta,\beta}$ of P_θ, i.e., $P_\theta(C_{S,\theta,\beta}) = \beta$.

Then, $\theta \in C_\beta(\mathbf{x})$ means that $S(\mathbf{x},\theta) \in C_{S,\theta,\beta}$. This indicates that P_θ fits $\mathbf{x}$ so well that $\mathbf{x}$ looks like a "realistic" outcome (at level β) if P_θ were the DGP.

Some observations:

- Note that this actually *does* refer to P_θ as *potential* DGP for $\mathbf{x}$. It is still "model-free" in the sense that $\mathbf{x}$ is treated as fixed and no DGP or "true θ" is assumed to have generated $\mathbf{x}$.

- This is probably somewhat hard to grasp for the non-expert.

- It does not involve infinite repetition of a real experiment.
- It also is a valid interpretation for many classical confidence intervals, namely those based on statistics S of the given kind.
- In earlier work Davies (2014) used the terms "adequacy region" or "approximation region". It seems to me that these capture the meaning somewhat better than "equivalence region".

REFERENCES

P. L. DAVIES (2014). *Data Analysis and Approximate Models*. Chapman and Hall/CRC, New York.

P. L. DAVIES (2024). *An approximation based theory of linear regression*. URL https://arxiv.org/abs/2402.09858.

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

C. HENNIG (2023). *Probability models in statistical data analysis: Uses, interpretations, frequentism-as-model*. In B. SRIRAMAN (ed.), *Handbook of the History and Philosophy of Mathematical Practice*, Springer, Cham, pp. 1411–1458. URL https://arxiv.org/abs/2007.05748.