ISSN 1973-2201 *STATISTICA, anno LXXXIV, n. 2, 2024*
DOI: https://doi.org/10.60923/issn.1973-2201/22203

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

Moritz Herrmann
LMU Munich, Munich Center for Machine Learning, Germany
Michael Herrmann [1]
Eberhard Karls University Tübingen, Germany

We congratulate Lutz Dümbgen & Laurie Davies for their new approach to measuring the relevance of covariates in linear regression (Dümbgen and Davies, 2024). The approach is based on a randomization scheme that replaces observed covariates with Gaussian covariates – i.e. covariates that are known to be irrelevant – and leads, inter alia, to the formulation of *Gaussian p-values* to measure the observed covariate's relevance without assuming a stochastic model for the given data. That means, their randomization scheme avoids any (implicit) reference to a "true" representation of data which is usually assumed to hold in statistical estimation problems. Yet, the obtained Gaussian p-values are still formally equivalent to model-based ones. This is achieved without resorting to asymptotic reasoning, distinguishing the approach from permutation tests, which are mathematically equivalent to exact tests only as the sample size goes to infinity, see Good (1994). We think this is a truly genuine and remarkable piece of work deserving wide recognition. Moreover, it needs emphasis that their approach to linear regression is based on a general approximation framework developed by Davies (2014) that avoids the assumption of "assumed (revealed?) truth" (Tukey, 1993, p. 2). This framework has not gained much attention yet and furthering it is important in our opinion as it appears to be universally applicable and to be a better account for the complexity of data-based inference compared to conventional statistical parlance. In the following, we want to comment on a connection to a fundamental issue not only of statistical research.

The Big Picture: Indoctrination, Cultural Divides, Statistical Reification
Davies (2024) sees parts of the statistical community affected by "indoctrination" because their papers on linear regression that build on the said (unconventional) approximation framework have been very difficult to get published. This assessment contains

[1] Corresponding Author. E-mail: michael.herrmann@tfw.uni-tuebingen.de

a valid and important criticism as it addresses a severe deficiency of current scientific practice: a widespread tendency to judge research mainly on the basis of whether it follows established conventions rather than its genuine scientific contribution. But it is also a harsh verdict because we think indoctrination implies and emphasizes bad intent of individuals. By contrast, we think of these issues as cultural divides that run through scientific communities. No one needs to have inherently bad intentions to follow or pass on beliefs, rituals, norms of a certain culture. Moreover, such conventions are often established unconsciously and remain unquestioned within the community, which makes them more difficult to overcome than if they were deliberately propagated.

We argue that in the statistical community one such divide centers around "statistical reification, the tendency to forget that mathematical arguments say nothing about reality except to the extent the assumptions they make (which are often implicit) can be mapped into reality in a way that makes them all correct simultaneously" (Greenland, 2023, p. 911). For example, there is the (unresolved) dispute between R. A. Fisher and J. Neyman over statistical testing. In particular, Fisher (1955, p. 69) opposed the *repeated sampling from the same population* assumption and averred: "Mathematicians without personal contact with the Natural Sciences have often been misled by such phrases. The errors to which they lead are not always only numerical." A second example are the two different cultures of statistical modeling as discussed by Breiman (2001, p. 202), who emphasized: "The belief in the infallibility of data models was almost religious. It is a strange phenomenon — once a model is made, then it becomes truth and the conclusions from it are infallible." Note, however, that Breiman's praise of the algorithmic modeling culture (and hence the devaluation of the data modeling culture) was mainly tied and limited to prediction problems.[1] In summary, that means one community prioritizes mathematical reasoning that provides conclusive and irrevocable truth. The other prioritizes empirical adequacy of conclusions drawn from these arguments, which implies accepting a degree of inconclusiveness, uncertainty, subjectivity that certainly is dubious for a formal way of thinking, but does certainly not imply justifications without rational argument.

A great merit of Dümbgen & Davies is that their approach is both formally precise and empirically adequate, and thus suitable for bridging this cultural divide. In particular, the suggested model-free approach to linear regression could be understood as being orthogonal to the dichotomy established by the two cultures of data modeling described by Breiman. Coincidentally, this work is also a cautionary example of how conventions can be a threat to diversity in perspective and how easily important, but unconventional research can fall victim to unquestioned beliefs and norms of an established scientific culture.

[1] This restriction was already critized in the discussion to Breiman's article by Sir David Cox and Bradley Efron (Breiman, 2001, pp. 216–219).

REFERENCES

L. BREIMAN (2001). *Statistical modeling: the two cultures (with comments and a rejoinder by the author).* Statistical Science, 16, no. 3, pp. 199–231.

P. L. DAVIES (2014). *Data Analysis and Approximate Models: Model Choice, Location-Scale, Analysis of Variance, Nonparametric Regression and Image Analysis*. CRC Press, New York.

P. L. DAVIES (2024). *Statistics: truth, ontology, approximation, honesty and indoctrination*. Statistica, 84, no. 2, pp. 83–92.

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

R. FISHER (1955). *Statistical methods and scientific induction*. Journal of the Royal Statistical Society: Series B (Methodological), 17, no. 1, pp. 69–78.

P. GOOD (1994). *Permutation Tests*. Springer Series in Statistics, New York.

S. GREENLAND (2023). *Connecting simple and precise P-values to complex and ambiguous realities (includes rejoinder to comments on "Divergence vs. decision P-values").* Scandinavian Journal of Statistics, 50, no. 3, pp. 899–914.

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.