ISSN 1973-2201 *STATISTICA, anno LXXXIV, n. 2, 2024*
DOI: https://doi.org/10.60923/issn.1973-2201/22201

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

Pietro Coretto[1]
Department of Economics and Statistics, University of Salerno, Italy

First, I want to congratulate Prof. Lutz Dümbgen and Prof. Laurie Davies for this interesting and fresh approach to linear modeling (Dümbgen and Davies, 2024). In statistical analysis, linear regression is a cornerstone for understanding relationships between variables. The authors explore the fascinating connection between two distinct approaches to linear regression: the model-based and the model-free perspective. In particular, the paper deals with the problem of assessing whether the response variable is related to the regressors or a subset of them by comparing the squared residuals obtained from the ordinary least squares (OLS) fit of the full model against those of the restricted model not including all the regressors. In the traditional model-based approach, based on the classical linear regression model, the previous comparison leads to the well-known F-test, for which exact p-values are available. The idea is to use the p-values to assess the relevance of the regressors. The authors propose a model-free framework that does not make assumptions about the underlying data-generating process. The surprising result is that the p-values derived for assessing the relevance of certain regressors in this model-free context coincide with those derived under the classical model-based setting. In addition to this key insight, the paper introduces the concept of "equivalence regions," which provides a new way to think about confidence regions in the model-free setting.

The idea that a "true" model does not exist and the quest for model-free methods is certainly not new in the literature. Usually, we label statistical methodologies that do not assume a specific distribution for some or all the variables involved as model-free. However, for the vast majority of methodologies labeled as model-free, we still make some (possibly weak) assumptions on the stochastic behavior of the observable quantities of interest (*e.g.* moment conditions, the existence of density functions, *etc.*). In my opinion, the most interesting aspect of this paper is that the authors push the model-free

[1] Corresponding Author. E-mail: pcoretto@unisa.it

approach into a new territory where data are assumed to be fixed (both the response and the regressors), whereas stochastic models are introduced to randomly perturb the data and define a test statistic that allows to measure the strength of a null hypothesis. My understanding is that the approach presented in this paper is strongly connected with the idea of "*data approximation*" developed in Davies (1995) (see also Davies, 2008, for further references). The author conceptualized this fixed-data paradigm, in which models are not true but possibly adequate if samples generated under the model are very similar to the samples actually obtained. Similarly, in this paper, the irrelevance of covariates is assessed by computing the probability that data randomly perturbed under a model consistent with the null hypothesis (irrelevance of covariates) produce a least square fit not worse than that produced by all covariates. In some sense, this is equivalent to assessing whether the null random model is an adequate representation of the observed data. I wonder whether a unifying theoretical framework that also includes the prediction problem is possible. In fact, in some fields of science, a feature is considered relevant if it has some predictive power for the response variable. The predictive paradigm of data science is dominant in today's applications. The question is how this fascinating construction that views observed data as fixed can cope and extend to frameworks where the method's performance on unseen data is central.

References

P. L. Davies (1995). *Data features*. Statistica Neerlandica, 49, no. 2, pp. 185–245. URL http://dx.doi.org/10.1111/j.1467-9574.1995.tb01464.x.

P. L. Davies (2008). *Approximating data*. Journal of the Korean Statistical Society, 37, no. 3, pp. 191–211.

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