Statistical Science
2019, Vol. 34, No. 4, 575–579
https://doi.org/10.1214/19-STS725
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Discussion of Models as Approximations I & II

Dag Tjøstheim

"All models are wrong but some are useful." This famous quote is attributed to George Box. The authors prefer to quote David Cox: "It does not seem helpful just to say all models are wrong. The very word model implies simplification and idealization."

The authors stress the model approximation aspect in their two interesting and inspirational papers. The first paper is concerned with linear regression models, or rather with regression functionals which are linear in the parameters, and where the functional itself is an OLS functional. In the second paper, more general regression functionals are treated, including likelihood-like functionals where nonlinearities can be meaningfully discussed.

The use of linear models as approximations is perfectly legitimate of course and is probably the most common approximation used in statistics, quite often with an additional Gaussian distributional assumption. In the first paper, the authors deviate from perhaps most contributors in that they try to find an interpretation of slope parameters as seen from the general viewpoint of a more correct and possibly nonlinear model. Moreover, they examine estimation errors under this perspective. Most users would be satisfied with evaluating these properties under the assumption that the linear model is correct.

The errors of parameter estimates in both papers are decomposed into two components; one component due to natural stochastic variations which may well be heterogeneous, and one component that is due to model errors. In the first paper, the authors use the ratio, the RAV, between a model trusting error and a model robust error to test model fit, and in the second paper they suggest that a well-specification test can be based on reweighting the data.

The model trusting error might be quite large of course, if, as in some cases in the first paper, the true model is strongly nonlinear, and the OLS regression functional by default ends up in a linear structure. The authors quote Freedman's somewhat provocative statement in this case where "... it is quite another thing to ignore bias [nonlinearity]. It remains unclear why applied workers should care about the variance of the estimator for the wrong parameter." I must admit that I have some sympathy with this statement, at least if it can be very easily detected that a linear model is completely wrong with resulting slope parameter being close to meaningless.

The authors themselves admit that a general interpretation of a linear regression parameter is "vexing," and I am not completely convinced by the authors attempt in Section 10 in the first paper. I find it not so easy to grasp. Parts of the difficulties are, in my opinion, that the authors force a linear structure on something that might be better, or to a better approximation, be modeled by a nonlinear or nonparametric approach, where a concrete and easy to understand interpretation of *local* slopes can be found.

In this respect I find the second paper, where nonlinear regression models are allowed, to be more satisfying. Actually, one might think that the linear regression functional of paper 1 could have been addressed as a special case of the set-up in paper 2.

I applaud the general set-up with population based regression functionals to define population parameters by extremal values of the functionals. In this sense the authors' approach is model free. The estimated parameters can then be obtained by minimizing the estimated regression functionals and consistency and asymptotic distributions follow quite straightforwardly. If one agrees that it is (almost) always meaningful to find errors of these estimates, the two-component decomposition of the errors is useful and ties in admirably with two basic papers by Hal White, where White (1980) is very much cited and deals with the purely random noise components, and White (1981) is much less cited and concerns the errors of model maladjustment.

The authors' approach makes for interesting and sometimes quite controversial reading for reasons

Dag Tjøstheim is Professor Emeritus, Department of Mathematics, University of Bergen, Norway (e-mail: Dag.Tjostheim@math.uib.no).