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).

mentioned already, one of them being the rather restrictive framework of a parametric linear approximation in paper 1. Nonparametrics are mentioned, especially in paper 2, where the authors also discuss additive models. In my view—and quite probably in the authors' view—in data analysis, it is not so much a question about a purely parametric model or a purely nonparametric approach, but more of an interplay between the two approaches. A parametric model or a parametric regression functional may benefit from a nonparametric exploratory analysis in the beginning and by a nonparametric diagnostic checking at the end of the data fitting process. Vice versa, a nonparametric approach may benefit from an underlying parametric structure made local resulting in local parametric regression functionals, as will be exemplified below.

In the further discussion, I have decided that rather than concentrating on details of the authors' approach, I will try to indicate possible extensions and complements of their work. I will highlight local parameter regression functionals, whose local parameters may have an easier interpretation than that attempted by the authors for the global linear OLS functional. I will also try to illustrate what kind of problems one might meet in an attempted extension to dependent data, both time series and spatial data, where difficulties of slope interpretation of the linear OLS functional may be exacerbated.

PARAMETRICS AND NONPARAMETRICS

Above I alluded to the interplay between nonparametrics and parametrics. If one is in doubt whether a given parametric structure, and in particular a linear one, is really appropriate for the data at hand, one may simply test for that parametric structure. There are many ways of doing this; see, for example, Härdle and Mammen (1993). The asymptotic theory of such functionals can be derived as in Härdle and Mammen (1993), but is not always accurate for a moderate sample size, so that bootstrapping can be recommended. In fact, Härdle and Mammen in their paper propose an exploratory test for a general parametric model $f(x,\theta)$ for the conditional mean, with a known f, measured against a nonparametric estimate of the same. They use the wild bootstrap for evaluating the corresponding test-of-fit functional.

In most cases, I believe that if there is an appreciable nonlinearity, this will be revealed by such an exploratory test. It will certainly immediately detect a nonlinearity of parabolic form such as $Y = a + bX^2 + \varepsilon$ mentioned as an example with a difficult-to-interpret slope in Section 10 of the authors' paper 1. One may argue that one may just go ahead fitting a linear model suggested by the OLS functional, and then test residuals in a goodness-of-fit test, of which there are numerous ones, and where the authors introduce a new one in terms of RAV. But one may have the impression that the authors in paper 1 also want to consider the properties of parameter estimates as if the linear model is an end product even in the face of strong evidence of nonlinearity.

LOCAL PARAMETER REGRESSION FUNCTIONALS

The introduction of regression functionals is not necessarily limited to the parametric case. Also nonparametric estimates such as the Nadaraya–Watson estimator for the conditional mean may be so introduced. Moreover, local linear regression functionals may be introduced resulting in local parameters.

First, if one considers the functional $E(Y - m(\mathbf{X}))^2$ for an unknown m, it is well known that the optimal solution, when it comes to minimizing it, is given by $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. This functional can be made local by introducing a kernel function such that one considers, (for simplicity of notation I consider the case of a scalar Y, X only),

$$G(x) = E\big[(Y-a)^2 K_h(X-x)\big],$$

where $K_h(\cdot) = K(\cdot/h)/h$ with K being a kernel function and h a corresponding bandwidth. Finding the minimum leads to the local parameter $a = a(x) = \frac{E(YK_h(X-x))}{E(K_h(X-x))}$ or by plug in, the Nadaraya–Watson estimator,

$$\hat{a}(x) = \frac{\sum_i Y_i K_h(X_i - x)}{\sum_i K_h(X_i - x)}$$

for given observations $X_1, \ldots, X_n$. This gives the local constant estimator of the conditional mean.

More appropriate in the context of the linear OLS regression functional of the authors is a local linear regression functional obtained by minimizing the local OLS functional

$$E\big[\big(Y - a - b(X-x)\big)^2 K_h(X-x)\big].$$

The estimated local regression parameters can then be found by considering and minimizing the corresponding empirical functional

$$\sum_i \big(Y_i - a - b(X_i - x)\big)^2 K_h(X_i - x)$$

leading to the estimate of the vector $[a, b]^T$,

$$[\hat{a}, \hat{b}]^T = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

and where $\mathbf{W} = \text{diag}\{K_h(X_i - x)\}$. This is the well-known local linear conditional mean estimator; see, for example, Fan and Gijbels (1996, Chapter 3). If again $m(x) = E(Y|X = x)$, then as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$, under weak regularity conditions $[\hat{a}, \hat{b}] = [\hat{a}(x), \hat{b}(x)]$ converges in probability to $[m(x), m'(x)]$, and asymptotic normality is obtained.

The slope parameter can be given a much more satisfying interpretation in the local linear case just presented. For a finite and moderate h, the slope parameter $b(x) = m'(x)$ measures approximately the change $m'(x)\Delta x$ in Y as X increases from x to $x + \Delta x$. As h and Δx decrease, the approximation is becoming more precise.

The authors treat vector regression models where $\mathbf{X}$ is a d-dimensional vector. In principle, the local linear analysis can be extended to this case, but in practice it is hit by the curse of dimensionality as d increases. An often used device for circumventing the curse is to assume that one might approximate the conditional expectation

$$m(\mathbf{x}) = E[Y|X_i = x_i, i = 1, \ldots, d]$$

by an additive form

$$m_0 + \sum_{i=1}^{d} m_i(x_i).$$

One seeks the optimal approximation in the least-squares sense, which is formulated as a nonlinear (additive) regression functional below. To make the terms identifiable, it is required that $\int m_j(y)p(y)\,dy = 0$, $j = 1, \ldots, d$, where p is the marginal density of Y. In general, one then obtains the optimal additive approximation by minimizing the functional

$$E\left[Y - m_0 - \sum_{i=1}^{d} m_i(X_i)\right]^2 \tag{1}$$

with

$$m_0 + \sum_{i=1}^{d} m_i(\cdot) \in \mathcal{F}_{\text{add}},$$

where $\mathcal{F}_{\text{add}}$ is the function space

$$\mathcal{F}_{\text{add}} = \Bigg\{ m_0 + \sum_{i=1}^{d} m_i(x_i) | m_0 \in \mathbb{R}, \int m_i(y) f(y)\,dy = 0 \text{ for } 1 \le i \le d \Bigg\}.$$

Mathematical details can be found in Mammen, Linton and Nielsen (1999). Estimates can subsequently be obtained by smooth backfitting. This is the algorithm for the analogue of the Nadaraya–Watson estimator (locally constant) implemented for the conditional expectation. For a general discussion of the smoothing backfitting algorithm, including the one based on more efficient local linear estimation I refer to Mammen, Linton and Nielsen (1999).

Again a local interpretation of slopes of the various components can be given.

Local parameter arguments are not limited to local least squares functionals. It is also possible to make parameters of general distributions local and use a local likelihood functional to define local population parameters $\boldsymbol{\theta}(\mathbf{x})$ in an approximating family $\{p(\cdot, \boldsymbol{\theta}(\mathbf{x}))\}$ of densities. Subsequently, local likelihood arguments are used to estimate these local parameters. Hjort and Jones (1996) argue for such an approach with applications to estimating a density f. I have been involved in this work in the special case of a Gaussian approximating family $\{\psi(\cdot, \boldsymbol{\theta}(\mathbf{x}))\}$. In that case, for a two-dimensional $\mathbf{x}$ a local population parameter $\boldsymbol{\theta}(\mathbf{x})$ can be determined by minimizing the functional

$$q = \int K_h(\mathbf{v} - \mathbf{x})\big[\psi\big(\mathbf{v}, \boldsymbol{\theta}(\mathbf{x})\big) - \log \psi\big(\mathbf{v}, \boldsymbol{\theta}(\mathbf{x})\big) f(\mathbf{v})\big]\,d\mathbf{v}.$$

Here, $\mathbf{v}$ is a two-dimensional running variable, and f is the density sought approximated. Further, $\psi(\cdot)$ is a bivariate normal distribution with local parameter vector $\boldsymbol{\theta}(\mathbf{x})$ defined by the 5 parameters of the bivariate normal distribution, namely the two means, the two variances and the correlation. These local population parameters can be estimated by a local likelihood method, with consistency and asymptotic normality obtained under regularity conditions. A number of applications, among others to a local correlation function, have been given. See, for example, Tjøstheim and Hufthammer (2013) and Lacal and Tjøstheim (2019) for references.

THE DEPENDENT CASE

The authors do not treat the dependent variable case. I am not going to treat this case extensively either, since that would require separate papers, which possibly the present authors are going to embark on. Here, I will just point out a few points which makes this extension nontrivial.

In the i.i.d. situation, the parabola case $Y = a + bX^2$ mentioned in Section 10 in the first paper may possibly be considered as contrived and easy to detect by any kind of nonparametric exploratory analysis. However, in the time series ARCH-GARCH case an analogous construction is both realistic and difficult to detect by a linear regression functional approach.

To take the simple example of a first-order ARCH model given by

$$Y_t = \sqrt{a + bY_{t-1}^2}\varepsilon_t,$$

where $\{Y_t\}$ is the observed time series and $\{\varepsilon_t\}$ is a series of hidden i.i.d. zero mean variables independent of $\{Y_t\}$ with variance σ_ε^2. In this case, which is also the case for the more general GARCH models, the correlation between Y_t and Y_s is zero for any $t \neq s$. The linear regression functional will in this case completely fail to reveal the (volatility) structure of the process, and this structure is very important in finance. And in this case it is not equally easy to detect the nonlinearity structure doing a local regression exploratory analysis; it will fail since $E(Y_t|Y_{t-1}) = 0$. A regression functional can be used on the process $\{Y_t^2\}$, or a more general likelihood type functional can be used as indicated in the second paper.

Another possible stumbling block in an extension to the dependent case is the authors' elegant use of bootstrap arguments to create a more stable estimator than the sandwich estimator. It is not immediately clear how this can be generalized to the time series case. It would be interesting to see if the block bootstrap or the stationary bootstrap could be used. It should be noted that Hal White (see Gonçalves and White (2004)) has done important work in this area as well. See also more recent work by Nordman and Lahiri (2014). A problem in the potential application to GARCH type processes is that in existing algorithms the block length seems to be determined by the autocorrelation function of the time series, and this is identically zero for nonzero lags for GARCH processes. One possible way out could be to use the existing formulas for block length on the squares of the process.

I will close this discussion with a few remarks on spatially dependent variables. It is well known that OLS estimation does not work well for spatial variables, at least not for so-called simultaneous autoregressive (SAR) models on a regular lattice. This was demonstrated already in the classic paper by Whittle (1954). Even for the process on the line

$$Y_l = aY_{l-1} + bY_{l+1} + \varepsilon_l,$$

he points out that minimization of the regression functional $E(Y_l - aY_{l-1} + bY_{l+1})^2$ leads to nonsensical results. This is due to the fact that $aY_{l-1} + bY_{l+1}$ and ε_l are dependent and that $E[Y_lY_{l-1}] = E[Y_lY_{l+1}]$. This is avoided in the additive functional (1) when adapted to space (cf. Lu et al. (2007)) where in the line case, say, in general $E[Y_l|Y_{l-1}] \neq E[Y_l|Y_{l+1}]$. This is also a reason that conditional models are more popular than simultaneous models in spatial analysis. Another way of avoiding the OLS-problems for SAR models is to use the likelihood function, often the likelihood constructed via the spectral distribution. This was demonstrated already in Whittle's paper, where he used spectral arguments to derive what has subsequently been named the Whittle likelihood in the spatial case. This leads to a far more computational demanding problem in finding estimates because of the evaluation of a complexly structured Jacobi determinant.

REFERENCES

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

GONÇALVES, S. and WHITE, H. (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *J. Econometrics* **119** 199–219. MR2041897

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947. MR1245774

HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647. MR1416653

LACAL, V. and TJØSTHEIM, D. (2019). Estimating and testing nonlinear local dependence between two time series. *J. Bus. Econom. Statist.* **37** 648–660. MR4016160

LU, Z., LUNDERVOLD, A., TJØSTHEIM, D. and YAO, Q. (2007). Exploring spatial nonlinearity using additive approximation. *Bernoulli* **13** 447–472. MR2331259

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

NORDMAN, D. J. and LAHIRI, S. N. (2014). Convergence rates of empirical block length selectors for block bootstrap. *Bernoulli* **20** 958–978. MR3178523

TJØSTHEIM, D. and HUFTHAMMER, K. O. (2013). Local Gaussian correlation: A new measure of dependence. *J. Econometrics* **172** 33–48. MR2997128

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027

WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.* **76** 419–433. MR0624344

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449. MR0067450