Statistical Science
2019, Vol. 34, No. 4, 566–568
https://doi.org/10.1214/19-STS722
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Discussion of Models as Approximations I & II

Sara van de Geer

Abstract. We discuss the papers "Models as Approximations" I & II, by A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zao and K. Zhang (Part I) and A. Buja, L. Brown, A. K. Kuchibhota, R. Berk, E. George and L. Zhao (Part II). We present a summary with some details for the generalized linear model.

Key words and phrases: Misspecification, sandwich formula.

1. PARAMETERS AS FUNCTIONALS

The authors have written a thought-provoking paper (Part I & II), which takes the issue that models are only approximations to a higher level, describing clearly the consequences and how to deal with model misspecification.

As the authors argue, it helps to view a parameter as some function of the distribution instead of thinking of the distribution as some function of a parameter. This is very much in line with Huber's view; see Huber (1967) where the sandwich formula for the asymptotic variance appears.

Let me recall Huber's results in the context of M-estimation. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be i.i.d. copies of a random variable $\mathbf{x} \in \mathcal{X}$ with distribution P. Let be given for $b \in \mathcal{B}$ be given loss functions $\rho_b : \mathcal{X} \to \mathbb{R}$. Suppose we observe the sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The M-estimator is now

$$\hat{\beta}_N := \arg\min_{b \in \mathcal{B}} \sum_{i=1}^{N} \rho_b(\mathbf{x}_i).$$

This is an estimator of the target

$$\beta := \arg\min_{b \in \mathcal{B}} \mathbb{E}\rho_b(\mathbf{x}).$$

Under some standard conditions $\hat{\beta}_N$ is a consistent and asymptotically linear estimator of β. The influence function is

$$IF(\cdot) = -\Lambda^{-1}\psi_\beta(\cdot),$$

where $\Lambda := \Lambda(\beta)$, and where, for $b \in \mathcal{B}$, the function ψ_b and the matrix $\Lambda(b)$ are defined as $\psi_b := \frac{\partial}{\partial b}\rho_b(\cdot)$ and $\Lambda(b) := \frac{\partial}{\partial b'}\mathbb{E}\psi_b(\mathbf{x})$, respectively.

Thus

$$\sqrt{N}(\hat{\beta}_N - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J),$$

where the asymptotic covariance matrix is

$$J = \Lambda^{-1} V \Lambda^{-1},$$

where $V := V(\beta)$ with, for $b \in \mathcal{B}$, the matrix $V(b)$ given by $V(b) := \mathbb{E}\psi_b(\mathbf{x})\psi_b'(\mathbf{x})$. This is the population version of the famous sandwich formula. The asymptotic variance can be estimated by

$$\hat{J}_N := \hat{\Lambda}_N^{-1} \hat{V}_N \hat{\Lambda}_N^{-1}, \tag{1.1}$$

where $\hat{\Lambda}_N = \hat{\Lambda}_N(\hat{\beta}_N)$, and $\hat{V}_N := \hat{V}_N(\hat{\beta}_N)$ with

$$\hat{\Lambda}_N(b) := \frac{\partial}{\partial b'} \sum_{i=1}^{N} \psi_b(\mathbf{x}_i)/N$$

and

$$\hat{V}_N(b) := \sum_{i=1}^{N} \psi_b(\mathbf{x}_i)\psi_b'(\mathbf{x}_i)/N \quad (b \in \mathcal{B}).$$

The estimate $\hat{\Lambda}_N$ also occurs when doing the Newton–Raphson algorithm, as $\hat{\beta}_N$ is a fixed point of the iterations

$$\hat{\beta}_{\text{new}} = \hat{\beta}_{\text{old}} - \hat{\Lambda}_N^{-1}(\hat{\beta}_{\text{old}}) \frac{1}{N} \sum_{i=1}^{N} \psi_{\hat{\beta}_{\text{old}}}(\mathbf{x}_i).$$

Thus

$$\|\hat{\beta}_{\text{new}} - \hat{\beta}_{\text{old}}\|_2^2 = \text{trace}\big(\Lambda_N^{-1}(\hat{\beta}_{\text{old}}) \hat{V}_N(\hat{\beta}^{\text{old}}) \Lambda_N^{-1}(\hat{\beta}_{\text{old}})\big).$$

Sara van de Geer is Full Professor, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland (e-mail: geer@stat.math.ethz.ch).

The trace of the sandwich formula can therefore be seen as a measure of the numerical stability.

The bootstrap also nicely picks up the sandwich formula.

Apart from technical regularity conditions, there are no model assumptions, that is, as the authors call it, M-estimation can be "assumption-lean."

For example, suppose one believes the model $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ holds true, with Θ a given subset of $\mathbb{R}^p$. Suppose densities $p_\vartheta := dP_\vartheta/d\mu$ exist with respect to a given dominating measure μ for all ϑ and let $\hat{\theta}_N$ be the maximum likelihood estimator

$$\hat{\theta}_N := \arg\max_{\vartheta\in\Theta} \sum_{i=1}^N \log p_\vartheta(\mathbf{x}_i).$$

The target is then

$$\theta := \arg\max_{\vartheta\in\Theta} \mathbb{E}\log p_\vartheta(\mathbf{x}).$$

If the model is well specified, we have $P = P_\theta \in \mathcal{P}$. Otherwise, P_θ is the best approximation of P in terms of Kullback–Leibler information. The sandwich formula holds with

$$V = \mathbb{E}\left[\frac{\dot{p}_\theta(\mathbf{x})\dot{p}'_\theta(\mathbf{x})}{p^2_\theta(\mathbf{x})}\right] = \Lambda + \mathbb{E}\left(\frac{\ddot{p}_\theta(\mathbf{x})}{p_\theta(\mathbf{x})}\right).$$

2. REGRESSION

The paper considers a regression framework, where there is an "input" variable $\mathbf{x}$ and a "response" variable $\mathbf{y}$. This leads to a very interesting connection between misspecification and causality. Let me summarize part of their framework as follows. First, replace in the above the variable $\mathbf{x}$ by the pair $(\mathbf{x}, \mathbf{y})$. The estimator is thus

$$\hat{\beta}_N := \arg\min_{b\in\mathcal{B}} \sum_{i=1}^N \rho_b(\mathbf{x}_i, \mathbf{y}_i)$$

and the target is

$$\beta := \arg\min_{b\in\mathcal{B}} \mathbb{E}\rho_b(\mathbf{x}, \mathbf{y})).$$

The authors clarify that, with a fixed design in mind, one may aim at a different target, namely

$$\beta(\mathbf{X}) = \arg\min_{b\in\mathcal{B}} \sum_{i=1}^N \mathbb{E}\big[\rho_b(\mathbf{x}_i, \mathbf{y}_i)|\mathbf{x}_i\big],$$

with

$$\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_N)'.$$

By fixing $\mathbf{X}$, applying Huber's arguments for independent, nonidentically distributed observations, and finally integrating out, one gets

$$\sqrt{N}\big(\hat{\beta}_N - \beta(\mathbf{X})\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \Lambda^{-1}V_1\Lambda^{-1}\big)$$

with

$$V_1 = V_1(\beta) := \mathbb{E}\mathrm{Cov}\big(\psi_\beta(\mathbf{x}, \mathbf{y})|\mathbf{x}\big).$$

We use here the notation $\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\mathbf{z}\mathbf{z}' - (\mathbb{E}\mathbf{z})(\mathbb{E}\mathbf{z})'$ for the covariance matrix of a random vector $\mathbf{z}$. We have then

$$\begin{aligned} V &= \mathrm{Cov}\big(\psi_\beta(\mathbf{x}, \mathbf{y})\big) \\ &= \mathbb{E}\,\mathrm{Cov}\big(\psi_\beta(\mathbf{x}, \mathbf{y})|\mathbf{x}\big) + \mathrm{Cov}\big(\mathbb{E}\big[\psi_\beta(\mathbf{x}, \mathbf{y})|\mathbf{x}\big]\big) \\ &= V_1 + V_2. \end{aligned}$$

If the model is well specified, $\mathbb{E}[\psi_\beta(\mathbf{x}, \mathbf{y})|\mathbf{x}] = 0$ so that the second term V_2 vanishes. In particular, in a generalized linear model, where $\mathbf{x} \in \mathbb{R}^p$ (a row-vector) and $\mathbf{y} \in \mathbb{R}$, the loss function is of the form

$$\rho_b(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x}b, \mathbf{y}).$$

Defining $\mu(\mathbf{x}) := \arg\min_{z\in\mathbb{R}} \mathbb{E}[\rho(z, \mathbf{y})|\mathbf{x}]$, we see that

$$V = \mathbb{E}\mathbf{x}'\mathbf{x}\big(\sigma^2(\mathbf{x}) + \eta^2(\mathbf{x})\big),$$

where $\sigma^2(\mathbf{x}) := \mathbb{E}[\dot{\rho}^2(\mu(\mathbf{x}), \mathbf{y})|\mathbf{x}]$ and $\eta^2(\mathbf{x}) := \mathbb{E}[(\dot{\rho}(\mathbf{x}\beta, \mathbf{y}) - \dot{\rho}(\mu(\mathbf{x}), \mathbf{y}))^2|\mathbf{x}]$. In the particular case of a one parameter family in canonical form, we have

$$\begin{aligned} &\rho(z, y) = yz - d(z), \quad \dot{\rho}(z, y) = y - \dot{d}(z), \\ &\sigma^2(\mathbf{x}) = \mathrm{Var}(\mathbf{y}|\mathbf{x}), \quad \eta^2(\mathbf{x}) = \big(\dot{d}\big(\mu(\mathbf{x})\big) - \dot{d}(\mathbf{x}\beta)\big)^2, \end{aligned}$$

and

$$V_1 = \mathbb{E}\mathbf{x}'\mathbf{x}\sigma^2(\mathbf{x}), \quad V_2 = \mathbb{E}\mathbf{x}'\mathbf{x}\eta^2(\mathbf{x}).$$

In this case, moreover,

$$\Lambda = \mathbb{E}\mathbf{x}'\mathbf{x}\ddot{d}(\mathbf{x}\beta).$$

The authors of the paper have highlighted a very useful view on regression models. There is indeed a strong relation with the "invariance" principle as explored in Peters, Bühlmann and Meinshausen (2016). In (over)simplified wording: if $\mathbf{x}$ is causal for $\mathbf{y}$, the formula for the distribution of $\mathbf{y}$ given $\mathbf{x}$ does not change if one changes (the distribution of) $\mathbf{x}$. In the present setup model, misspecification means that the parameter changes if $\mathbf{X}$ changes. The authors propose to artificially change (the distribution of) $\mathbf{X}$ by using a reweighting scheme. This can be a very useful diagnostic tool for misspecification. The exact interpretation per coefficient is however less clear.

The idea of ancillarity is related, but somehow different. In regression and in simple words, ancillarity means that the distribution of $\mathbf{x}$ does not depend on the parameters of the model for $\mathbf{y}$ given $\mathbf{x}$. Thus $p(x)$ and $p(y|x)$ do not share parameters. On the other hand, suppose they do:

$$p_\beta(x, y) = p_\beta(y|x)p_\beta(x).$$

Then the one has for the Fisher information

$$I := I_1 + I_2,$$

where

$$I := \mathrm{Cov}\big(s_\beta(\mathbf{x}, \mathbf{y})\big), \quad I_1 := \mathbb{E}\,\mathrm{Cov}\big[s_\beta(\mathbf{y}|\mathbf{x})|\mathbf{x}\big],$$

and

$$I_2 := \mathrm{Cov}\big(\mathbb{E}\big[s_\beta(\mathbf{x}, \mathbf{y})|\mathbf{x}\big]\big) = \mathrm{Cov}\big(s_\beta(\mathbf{x})\big)$$

with

$$s_\beta(x, y) = \frac{\dot{p}_\beta(x, y)}{p_\beta(x, y)}, \quad s_\beta(y|x) = \frac{\dot{p}_\beta(y|x)}{p_\beta(y|x)},$$
$$s_\beta(x) := \frac{\dot{p}_\beta(x)}{p_\beta(x)}.$$

In other words, $\mathbf{x}$ contains information about the parameter which results in an increase in Fisher information. The situation in this paper is exactly the other way around. as formulated at the end of Section 4.1. It is the regressor *distribution* that affects the parameters (not the regressor itself) that is, $\beta(\mathbf{X})$ is a function of $\mathbf{X}$.

Finally, I think the idea to view parameters as functional of the distribution, instead of distributions determined by parameters, is very appropriate but may lead to less interpretability of these parameters. The methodology in the paper allows one to infer to a certain extent whether or not the model is misspecified. But even if the conclusion is that model misspecification is likely, I would take a pragmatic point of view and still interpret the estimates as if the model were correct.

REFERENCES

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, Calif. MR0216620

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186