Practical issues concerning assumption-lean inference for generalized linear models

Elizabeth L. Ogburn[1], Junhui Cai[2], Arun K. Kuchibhotla[3], Richard A. Berk[4], and Andreas Buja[2,5]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; email: *eogburn@jhsph.edu*
[2]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
[4]Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA
[5]Flatiron Institute, New York, NY, USA
The authors contributed equally to this comment.

August 14, 2021

We congratulate the authors on their excellent article (Vansteelandt and Dukes, 2021). In this comment we highlight a few practical issues related to their proposal.

Not all conditional associations between outcomes and exposures are of interest. Those that are tend to be directional: up or down. The simplest way to assess directionality is to fit a confounder-adjusted linear exposure term, as the authors propose. We agree with this approach as some of us have argued that linear slopes are meaningful and interpretable even if the directional association is not linear (Buja et al., 2019, Section 10). The authors, and Whitney et al. (2019), remind us that severely misspecified adjustment can result in distortions of linear exposure slopes. In their examples the $A - L$ distributions have U-shaped nonlinearities and, as a result, naive linear adjustment produces a biased estimate of the true slope. Thorough data analysis could unearth such exposure-confounder structure if present in real data. A greater worry for practitioners is missing an essential confounder that biases or reverses the direction of association. The authors' inferential framework does not require L to control for all $A - Y$ confounding, but meaningful use of the estimand likely does—and therefore practitioners should select L with care and interpret estimates in conjunction with sensitivity analyses.

The authors' project of assumption lean inference rests on the assumption that nuisance parameters can be estimated nonparametrically at rate $n^{1/4}$. It is surprising to us that this property is widely assumed to hold for machine learning methods. The authors are in good company with this assumption, but, for example, the random forests included in the authors' analyses can have large bias if a tuning parameter is chosen badly (Olson, 2018), and as far as we know cross-validation has

not been shown to reliably choose good tuning parameters. Even if $n^{1/4}$ rates are achieved asymptotically, slower rates of convergence may require large samples before asymptotic approximations are useful. This points to the importance of methods to test or help ensure that the required rates are achieved (Liu et al., 2020; Robins et al., 2008; van der Laan et al., 2021), or to perform valid inference under slower rates (Cattaneo and Jansson, 2018; Kuchibhotla et al., 2021).

We re-ran the authors' code and applied *HulC*, a new method for the construction of assumption-lean confidence intervals (Kuchibhotla et al., 2021)[1]. We found that the point estimates are indeed sensitive to choice of tuning parameters. Although HulC intervals are wider, they are valid even if approximate normality does not hold, as would be the case if the nuisance estimators converge slower than $n^{-1/4}$, as long as the estimator satisfies a weaker *median unbiasedness* property (Kuchibhotla et al. 2021).

Table 1: Comparison of confidence intervals calculated using the method proposed by Vansteelandt and Dukes (V&D) with those calculated using HulC. The effects of participating in First Steps and of maternal age on the continuous outcome *birth weight* were estimated using the Vansteelandt and Dukes estimator with a linear link function. The effects of participating in First Steps and of maternal age on the binary outcome *low birth weight* were estimated using the Vansteelandt and Dukes estimator with a logit link function. All the nuisance functionals were estimated using SuperLearner with the same library and covariates as in V&D (with the exception of support vector machines; see footnote). The estimates were sensitive to the choice of tuning parameters and to methods included in the SuperLearner library. The effect of age on low birth weight is significant at 0.05 level using the authors' proposed method but not using HulC.

	Dependent variable:			
	Birth weight		Low birth weight	
	V&D	HulC	V&D	HulC
First Steps	−7.1	−7.1	0.04	0.04
	(−85.1, 70.9)	(−159.4, 144.2)	(−0.44, 0.52)	(−1.60, 1.33)
Age	−1.6	−1.6	0.06	0.06
	(−6.5, 3.4)	(−14.0, 5.0)	(0.03, 0.08)	(−0.06, 0.16)

References

Buja, A., L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science 34*(4), 523–544.

[1]The data and code were provided by the authors. We modified the code slightly, removing the *support vector machine* method from the SuperLearner library because of an error message. Because of this, our point estimates are close, but not identical, to those reported by the authors. The code to produce all tables is available at `https://github.com/cccfran/HulC-on-VD`.

Cattaneo, M. D. and M. Jansson (2018). Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica 86*(3), 955–995.

Kuchibhotla, A. K., S. Balakrishnan, and L. Wasserman (2021). The HulC: Confidence regions from convex hulls. *arXiv preprint arXiv:2105.14577*.

Liu, L., R. Mukherjee, and J. M. Robins (2020). On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science 35*(3), 518–539.

Olson, M. (2018). *Essays on random forest ensembles*. Ph. D. thesis, University of Pennsylvania.

Robins, J., L. Li, E. Tchetgen, and A. van der Vaart (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics.

van der Laan, M., Z. Wang, and L. van der Laan (2021). Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*.

Vansteelandt, S. and O. Dukes (2021). Assumption-lean inference for generalised linear model parameters. *JRSS-B*.

Whitney, D., A. Shojaie, and M. Carone (2019). Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics 34*(4), 591.