Statistical Science
2019, Vol. 34, No. 4, 580–583
https://doi.org/10.1214/19-STS726
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Comment: "Models as Approximations I: Consequences Illustrated with Linear Regression" by A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, L. Zhan and K. Zhang

Roderick J. Little

1. OVERVIEW

I congratulate Buja et al. on this ambitious and detailed description of a vitally important topic in statistics. The question of how to account for modeling uncertainty is a fundamental problem of statistical inference. I found the Buja et al. papers both challenging and thought-provoking, and I appreciate the opportunity to participate in the discussion. I focus my remarks on the first paper, since the second one largely concerns generalizations that are not the focus of my remarks.

Buja et al. adopt a traditional frequentist perspective. In contrast, I approach the topic from a "calibrated Bayesian" philosophy of statistical inference, where the inference for a particular dataset is Bayesian, but models are chosen to attempt to achieve good frequentist operating statistics (Box, 1980, Rubin, 1984, 2019, Little, 2006, 2011). I also comment on two aspects that receive little attention in the Buja et al. papers, the role of the selection mechanism in statistical modeling, and the perspective of finite population sampling. In the modeling approach to finite population inference, the finite population is assumed to be sampled from an underlying infinite "superpopulation," so what Buja et al. call the "population" I will call the "superpopulation." As an advocate of the calibrated Bayesian approach to survey sampling (Little, 2004, 2012), the topic of Buja et al. is pertinent because, as they note, the Bayesian approach is fundamentally "model-trusting," whereas the competing design-based approach to survey inference is "model-skeptical" and "assumption-lean."

In support of the calibrated Bayes position, I contrast the Buja et al. papers to Szpiro, Rice and Lumley (2010), henceforth SRL, an excellent paper that provides a justification of sandwich estimation of standard errors from a Bayesian perspective. Buja et al. reference SRL, but do not compare it with their work.

Roderick J. Little is Distinguished University Professor, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, USA (e-mail: rlittle@umich.edu).

2. SIMPLICITY, NOT MATHEMATISTRY

The Buja et al. papers seem to me quite mathematically formidable, despite the absence of formal regularity conditions. The approach to relaxing assumptions seems to me abstract—I am not looking forward to attempting to explain to practitioners, struggling with the interpretation of a regression coefficient in a logistic regression, that the target slopes are actually projections on a nonparametric space. I argued in my Fisher lecture (Little, 2013) that a primary advantage of the Bayesian approach to statistics is its conceptual simplicity. If, like me, you find the level of mathematical sophistication in the Buja et al. papers challenging, I recommend the fundamental simplicity of the Bayesian perspective in SRL. That is not to say it is easy to implement, but the difficulties lie in developing an appropriate Bayesian model that captures the important scientific aspects of a problem without unnecessary "clutter." This is the "art" of statistics, and it distinguishes it from the field of mathematics.

3. TERMINOLOGICAL TORTURE: "RANDOM" VS. "FIXED" EFFECTS, AND "NONLINEARITY"

I have never resonated with the frequentist interpretation of what is "random" and what is "fixed." Effects in analysis of variance are called "random" if they are regarded as sampled from a population, and "fixed" if they are not; in Buja et al., "fixed" regressors become "random" under potential model misspecification. If X is a treatment indicator, in what sense is it "random"?

Whether randomly assigned or not, it is a "fixed" entity, and it is not clear to me why treating it as random is relevant for defining a treatment effect. As an aside, a treatment effect in a randomized clinical trial can be estimated by regression with fixed treatment indicator under minimal assumptions, so the "model-trusting" paradigm seems fine in that setting.

One of the reasons I gravitated to Bayes is that the Bayesian usage of "fixed" and "random" is much more straightforward and intuitive—a quantity is fixed if it is known, and random (in the sense of being assigned a distribution to quantify uncertainty) if unknown. From the Bayesian perspective, known factors in an analysis of variance are fixed, whether or not they can be regarded as sampled from a population. A "random effects" model assigns a proper prior on group means, leading to "borrowing of strength," and a "fixed effects" model assigns a flat prior distribution on the means, implying that no borrowing of strength is warranted.

Another term I find problematic is "nonlinear," as in "the sandwich estimator of the standard error ... is asymptotically correct even in the presence of nonlinearity." It is confusing enough that nonlinear can mean nonlinear in the parameters or the variables, but it adds to the confusion to use the term for model misspecification, given that nonlinear models are standard in some areas, such as pharmacokinetic/pharmacodynamic modeling. The authors' term "first-order model misspecification" in footnote 2 is not ideal but is a big improvement.

4. ROBUST CALIBRATED BAYES, VIA TARGET AND WORKING MODELS

The elements of Bayesian inference are (a) defining the estimand; (b) formulating a useful working model; (c) computing the posterior distribution of the estimand under the model; and (d) diagnostic checks and sensitivity analysis to assess the model and assess the performance of the inference. As SRL make clear, the estimand might be a function of the parameters in the model, rather than one of those parameters.

In a finite population setting, Little (2004, 2012) distinguishes a "target model" that determines the population quantities of interest, and a "working model" that is the basis for inference, and is used to predict survey variables for the nonsampled and nonresponding units in the population. The estimand is the population quantity obtained from fitting the target model to the entire population, using some agreed fitting principle such as maximum likelihood. For example, if the target model assumes that the mean of Y has a linear regression on X, then a target finite population quantity might be the slope of Y on X fitted to the entire population by least squares. This quantity exists regardless of whether the regression of Y on X is really linear, although its utility for summarization is weakened if the regression is highly nonlinear. The target finite population quantity is a useful target for inference even if the main interest is in "analytic" inference for the corresponding superpopulation parameter, because a poor estimate of the former is also a poor estimate of the latter.

For robust inference, the working model does not need to assume a linear regression of Y on X; minimally, I would argue that it needs to incorporate survey design features that are not necessarily part of the target model. One situation where this is important is in the context of probability samples where the selection probabilities are not constant, leading to sampling weights defined as the inverse of the probability of selection. The role of sampling weights in regression is a controversial topic, with social scientists often ignoring them and survey samplers using them to weight the units. My own view is that sampling weights need to be incorporated into the working model, though as predictors of nonsampled values rather than as weights for the sampled units. A flexible specification of the regression on the weights provides inference with a form of double robustness. This idea is pursued in the context of sampling weights in Zheng and Little (2005) and in the context of nonresponse weights in Little and An (2004) and Zhang and Little (2009).

SRL apply a similar approach to target and working models in their superpopulation regression setting. Specifically, their working model for a regression of Y on X is normal with a mean $\phi(x)$ and variance $\sigma^2(x)$ that are highly "nonparametric," that is do not make strong assumptions about the form of the regression function; perhaps "many-parametric" is a better term. They define the target parameter based on the least squares fit of the simpler target model—the linear regression of Y on X—to the superpopulation, as in their equation (3). This approach is a bit more abstract than my approach for a finite population, but still seems to me a lot simpler than the Buja et al. definition of slopes as projections on a non-parametric space.

Since Bayes inference is conditioned on the choice of working model, the Bayesian approach to "model-robust" inference is to make the model robust, rather than changing the interpretation of the estimand as in Buja et al. This could involve a flexible mean and variance function in the regression model, as in SRL's equation (3), or more flexible distributional assumptions, for example, replacing normality of errors by a longer-tailed distribution, or a Dirichlet process prior. The degree of complexity of the model should depend on the sample size, with simpler models being preferred for small sample sizes, and more complex ones for large sample sizes. The asymptotic considerations that dominate frequentist arguments have a role, but are limited by the fact that we don't know how large asymptotic is in many real applications.

5. CONDITIONING AND ANCILLARITY

Buja et al. emphasize the lack of ancillarity of X when the regression model is misspecified. I agree (isn't it obvious?) that arguments about whether statistics are sufficient or ancillary assume the validity of the model; the "Achilles heel" of otherwise impressive arguments in favor of the likelihood principle (Birnbaum, 1962) is that we don't know the true model, and working models are all to some degree misspecified. However, from my calibrated Bayesian perspective, ancillarity is irrelevant for the inference itself, because the posterior distribution conditions on all the data; questions of ancillarity are secondary, in that they relate to the reference set for assessing operating characteristics of the inference, but not the inference itself.

Concerning conditioning, one of the virtues of Bayes is that it emphasizes clarity about what is being conditioned in a probability statement (e.g., Little, 2013). The estimand either conditions on X as, for example, when prediction given X is the objective, or it does not, as in equation (3) of SRL. The posterior distribution of the estimand conditions on X, because it conditions on all the data.

6. THE ROLE OF THE SELECTION MECHANISM

Rubin (1974, 1976, 1978) highlights the importance of the sample selection mechanism in the robustness of model inferences by incorporating the mechanism as part of the model. Buja et al. state a key random sampling assumption (italics mine):

> "In fact, it may rely on *no more* than the assumption that the rows (y_i, x_i) of the data matrix … are i.i.d. samples from a joint multivariate distribution subject to some technical conditions."

I add italics because this assumption, routine in much of mathematical statistics, is both crucial—in some respects it trumps all the other assumptions in the statistical model—and often very questionable. If, as is usual, units are not selected by simple random sampling, this is an assumption, and if violated then estimates under model-trusting or model-robust paradigms—are subject to unknown biases. Units are rarely selected by random sampling from a population, as with "found data" not subject to a statistical design for data collection, or clinical trials where participants are volunteers, not randomly selected from the target population for a drug. In finite population sampling, the assumption of a simple random sample is extreme and rare, because of limitations in the sampling frame or nonresponse. Concerning nonresponse, the random sampling assumption translates to missing completely at random (Rubin, 1976), an assumption that is rarely satisfied.

The conditioning on x's in the "model-trusting" paradigm requires only simple random sampling of the y's given the x's, which is a weaker assumption since it does not require that the x's are themselves randomly sampled. So, arguably the "model-robust" approach trades weakening the model assumptions for strengthening the random sampling assumption, which is often highly questionable.

7. CONCLUSION

To conclude with a point of agreement, the authors make a good case for using the estimate of the RAV as a diagnostic for potential model misspecification. This usage accords with my "calibrated Bayesian" philosophy, which allows for frequentist notions of model checking (Rubin, 1984); since the two variance estimates are based on contrasting perspectives on model misspecification, comparing them seems a good idea.

I applaud Buja et al. for thought-provoking work on a topic of great importance, even though ultimately I cling to my Bayesian "model-trusting" paradigm, albeit with flexible, well-calibrated working models that can earn my trust.

REFERENCES

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326. MR0138176

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

LITTLE, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Amer. Statist. Assoc.* **99** 546–556. MR2109316

LITTLE, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* **60** 213–223. MR2246754

LITTLE, R. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statist. Sci.* **26** 162–174. MR2858391

LITTLE, R. J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *J. Off. Stat.* **28** 309–337.

LITTLE, R. J. (2013). In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. *J. Amer. Statist. Assoc.* **108** 359–369. MR3174626

LITTLE, R. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

RUBIN, D. B. (2019). Conditional calibration and the sage statistician. *Surv. Methodol.* To appear.

SZPIRO, A. A., RICE, K. M. and LUMLEY, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Ann. Appl. Stat.* **4** 2099–2113. MR2829948

ZHANG, G. and LITTLE, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* **65** 911–918. MR2649864

ZHENG, H. and LITTLE, R. J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *J. Off. Stat.* **21** 1–20.