Statistical Science
2019, Vol. 34, No. 4, 569–571
https://doi.org/10.1214/19-STS723
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Comment on Models as Approximations, Parts I and II, by Buja et al.

Jerald F. Lawless

Abstract. I comment on the papers Models as Approximations I and II, by A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao and K. Zhang.

Key words and phrases: Covariate distributions, misspecification, regression models, transportability.

Buja et al. provide an interesting and valuable discussion of certain aspects of regression methodology. They deal with the common setting where covariates (regressors) $\mathbf{X}$ are random, and not fixed values assigned as part of an experimental plan. A central theme in the papers is that standard model-based inference procedures assume the correctness of a family of models for the conditional distribution of a response variable Y, given $\mathbf{X}$, but models never represent exactly the true distribution. For the sake of discussion, I will refer to this as model misspecification. In Part I, the authors focus on the conditional mean function $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ and models of the form $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$; more generally one can consider models $F(y|\mathbf{x};\boldsymbol{\theta})$ for the conditional distribution function of Y, indexed by a parameter $\boldsymbol{\theta}$. They consider the consequences of model misspecification, which are summarized in points (1) to (10) at the start of Section 14 in Part I. The fact that models are approximations to reality, and the consequences of misspecification, are widely known and recognized in good statistical practice but I agree with the authors that they are given insufficient attention in most teaching of statistics.

I'll begin with brief comments on the mathematical results central to the paper, with some rephrasing and standard notation. Suppose a (working) family of models is indexed by a parameter $\boldsymbol{\theta}$ and that $\hat{\boldsymbol{\theta}}_n$ denotes an estimator based on a random sample of n pairs $(Y, \mathbf{X})$ that is consistent for $\boldsymbol{\theta}$ when the model is correct. It is well known that when the assumed model is not correct, $\hat{\boldsymbol{\theta}}_n$ converges in probability to some $\boldsymbol{\theta}^*$ as $n \to \infty$ under mild conditions. Buja et al. focus on the fact that $\boldsymbol{\theta}^*$ is a functional $\boldsymbol{\theta}(P_{YX})$ of the true distribution P_{YX} of $(Y, \mathbf{X})$. They derive and use versions of the well-known result that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically normal with mean zero and covariance matrix of the form $\boldsymbol{\Sigma} = A^{-1}(\boldsymbol{\theta}^*)B(\boldsymbol{\theta}^*)A^{-T}(\boldsymbol{\theta}^*)$. In the regression setting of Part I White's variance estimator (2), though first introduced by him as a "heteroscedasticity-consistent" estimator, is noted to be consistent for $\boldsymbol{\Sigma}$ more generally. They term such estimators "model-robust" and stress that "model-trusting" variance estimators, for example, based on $A^{-1}(\boldsymbol{\theta}^*)$ for maximum likelihood estimators, should not be used. I would make two points here. First, robust variance estimators like (2) are widely used and not as obscure as the authors suggest: they can be obtained by using the fact if $\psi_i(\boldsymbol{\theta}) = \psi(Y_i, \mathbf{X}_i; \boldsymbol{\theta})$ denotes an estimating function component for observation i and $A_i(\boldsymbol{\theta})$ the corresponding negative Jacobian matrix (see Part II, Section 7.2), then since the $\psi_i(\boldsymbol{\theta}^*)$ are i.i.d., the component matrices in $\boldsymbol{\Sigma}$ can be estimated as sample averages with $n^{-1}\sum_{i=1}^n A_i(\hat{\boldsymbol{\theta}}_n)$ for $A(\boldsymbol{\theta}^*)$ and $n^{-1}\sum_{i=1}^n \psi_i(\hat{\boldsymbol{\theta}}_n)\psi_i(\hat{\boldsymbol{\theta}}_n)^T$ for $B(\boldsymbol{\theta}^*)$. This gives (2) for the OLS estimating function. Second, "model-trusting" variance estimates can be used in the random X setting; they should however be used cautiously in situations where one has sufficient confidence in the working model.

Buja et al. explore some of the consequences of misspecified models. In fact, they want to avoid models, though the parameter estimates are based on a specification of some kind. Points they emphasize are that one should focus on $\boldsymbol{\theta}(P_{YX})$ and that the distribution of $\mathbf{X}$ is an important consideration in doing this.

Jerald F. Lawless is Distinguished Professor Emeritus, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (e-mail: jlawless@uwaterloo.ca).

I agree about the importance of P_X and also with their points on the usefulness of comparing robust and model-trusting variance estimates. I also agree with their emphasis of the x–y bootstrap. The detailed analysis of the effects of nonlinearity on estimates from linear models, and on regression "error" and residuals, is useful in understanding effects of nonlinearity and also the behavior of specification tests based on comparison of robust and model-trusting covariance matrices. Part II has a number of ideas that are valuable at least qualitatively. One is the definition of a well-specified functional which, as they mention, has important connections to causal inference and the integration of information. An idea with more obvious quantitative application is the examination of the effects of alternative distributions for $\mathbf{X}$, using reweighting or otherwise. I'll come back to some of these points as I discuss a few important issues for regression analysis.

Buja et al. wish to avoid models, and ask what estimates represent when no model is assumed. The answer that they are functionals of the (true) data distribution is a familiar one in nonparametric statistics, but there the focus is on interpretable functionals such as means, standard deviations or quantiles. In the present setting, the crucial question is what $\boldsymbol{\theta}^* = \boldsymbol{\theta}(P_{YX})$ and related estimators represent when models that produce an estimator are misspecified. "Assumption-lean" approaches have many applications, but do not obviate the importance and usefulness of models. Modeling plays an important role in scientific discovery and progress. Buja et al. comment in Section 9 of Part II that "Regression is the attempt to describe the conditional response distribution $P_{Y|X}$." Such descriptions are most useful when they involve models that can be applied and compared across studies or populations. Assumptions about the form of certain distributions or processes, guided by background knowledge, help in deciding what features to examine and in developing scientific interpretations. Models need not be viewed as "true" to be useful; they are approximations to reality and careful checks on the adequacy of models is a key aspect of good statistical practice. Models provide best approximations in a certain sense and Buja et al. note that the definition of best involves the distribution of $\mathbf{X}$. Indeed, since the usefulness of a model in a specific setting will depend on the range and distribution of covariates, a definition without this property would not be very helpful. The roles of models have been extensively discussed and I will not repeat additional arguments here, but in addition to references in the paper see, for example, Cox (1990), Breiman (2001) and discussants, Shmueli (2010) and references in these papers.

Discussions of statistical methodology should be framed according to the objectives. A primary distinction is between situations where the objectives are scientific discovery or understanding, and situations where the objective is automated decision making. In the latter case, some procedures use predictive models but may not be concerned about interpretability, and others may employ algorithms that do not use a model. Some of the disagreements in the literature concerning the use of models (e.g., see Breiman, 2001 and discussants) arise because objectives are not clearly stated. Shmueli (2010) has recently discussed this; he refers to explanatory versus predictive objectives, noting that in some settings a good predictive model may not be transparent or easily interpreted. Buja et al. do not discuss this directly, but their points on interpretability, misspecification and "well specification" mainly address scientific understanding.

Let us suppose that we are interested in the conditional mean function $E(Y|\mathbf{x}) = \mu(\mathbf{x})$ and perhaps also the conditional standard deviation function $\sigma(\mathbf{x})$. It is crucial in considering these functions of $\mathbf{x}$ that we think carefully about the range and distribution of $\mathbf{X}$ in the population of interest, and relationships between individual covariates X_j in $\mathbf{X}$. As the paper stresses, this is important for the interpretation of parameter estimates and for assessing the adequacy of models in specific settings. It hardly needs saying, for example, that a linear model may be adequate over a limited range for $\mathbf{x}$ but be quite unsatisfactory over a broader range. Other things that guide how we study $\mu(\mathbf{x})$ include rough ideas about standard deviation to mean ratios $\sigma(\mathbf{x})$: $\mu(\mathbf{x})$, scientific knowledge, and the type and amount of data available. Exploratory data analysis and alternating bouts of model fitting and assessment are used in many settings, and create difficulties for formal inferences concerning covariate effects and "final" models.

The papers aim to develop a model-free theory for regression, but do not say much about how regression analysis should be done. They do however consider two tools for understanding relationships between Y and covariates, based on RAV ratios of robust to model-trusting standard errors for individual regression coefficients β_j and, in Part II, on covariate reweighting. These seem useful for model assessment, but it would be good to see comparisons on real data with other techniques, including residual analysis, nonparametric or weakly parametric estimation of $\mu(\mathbf{x})$ and calibration checks. Comprehensive model assessment is difficult in settings involving many covariates, and testing and sensitivity analysis based on model expansion are key tools. Assessment of prediction error for models is useful even if our main objectives are scientific understanding and explanation; models with small prediction error engender confidence in the relevance and interpretation of parameter estimates. The authors prefer to avoid models and discuss how estimates $\hat{\beta}_j$ for individual covariates can indicate directions of association for responses and adjusted covariates, though association here means linear association. The connection with weighted averages of case-wise slopes will sometimes be helpful but exactly when seems unclear in the absence of more information about the nature of $\mu(\mathbf{x})$. The relationship between a covariate and Y is affected by associational and causal relationships among the covariates, and difficulties in assessing the importance of a covariate are well known. Shmueli (2010) and others argue that different approaches to modeling and assessment of importance are needed for explanatory versus predictive purposes. For scientific understanding, qualitative assessments using paradigms of causal inference or conditional independence offer some help, but do not in themselves look at strengths of associations or effects. When the objective is understanding, I find it difficult to envision truly comprehensive analysis without considering both response models and covariate distributions.

The importance of the **X**-distribution for causal analysis is noted in Part II, Section 3.4. A related discussion concerns the transportability of inferences from one population or group to another. Numerous authors have noted that inferences or models for conditional distributions such as $F(y|\mathbf{x})$ are generally more transportable than inferences about marginal effects or distributions; for example, see Keiding and Louis (2016) and discussion of the paper. In Buja et al.'s terms, "well-specified" functionals and their estimates are more transportable than functionals that are not well specified. A major reason for differences in transportability is that marginal distributions or distributions that condition on just a specified covariate involve averages over the covariate distribution; these distributions routinely vary from population to population, even if the conditional distributions are (approximately) the same. This is a factor in the failure of results from both randomized intervention studies and observational studies to hold up in other settings. It also makes the comparison of "causal" effects across different studies challenging; the same issue applies to meta-analysis. Still another problem occurs with attempts to use data from an external data base or study to augment (improve estimation in) the analysis of a specific study. Chatterjee et al. (2016) and Han and Lawless (2019) show that differences in the covariate distributions in the specific study and the external data can result in substantial bias in the augmented estimates of regression coefficients. Several of Buja et al.'s points are important here and, in particular, the need for careful consideration of the **X**-distribution.

Model misspecification is a key source of erroneous inferences and of failures in prediction and decision-making, and this paper is a welcome contribution. It seems to me that the insights from the authors' model-free theory are also found using standard results concerning working models and estimation theory under model misspecification. Nevertheless, I grant the authors their point of view, and thank them for drawing attention to the importance of model misspecification and to the distribution P_X in regression settings.

REFERENCES

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. MR1874152

CHATTERJEE, N., CHEN, Y.-H., MAAS, P. and CARROLL, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.* **111** 107–117. MR3494641

COX, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5** 169–174. MR1062575

HAN, P. and LAWLESS, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statist. Sinica* **29** 1321–1342. MR3932520

KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587

SHMUELI, G. (2010). To explain or to predict? *Statist. Sci.* **25** 289–310. MR2791669