Statistical Science
2019, Vol. 34, No. 4, 572–574
https://doi.org/10.1214/19-STS724
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Comment: Models as Approximations

Nikki L. B. Freeman, Xiaotong Jiang, Owen E. Leete, Daniel J. Luckett, Teeranan Pokaprakarn and Michael R. Kosorok

1. INTRODUCTION

We congratulate Andreas Buja and his coauthors on their thought provoking and ambitious work, "Models as Approximations, Parts I and II." This work deeply examines the meaning of model robustness, the consequences of model misspecification and culminates in the formulation and development of the notion of "well-specified" regression. Although the regressors-as-fixed point of view of regression has dominated statistical practice, the work of Buja et al., adds to a growing literature on the implications of random regressors and model misspecification on inference and prediction. We do not endeavor to nor intend to enumerate those here but will mention a few to give a sense of the literature. For example, Sen and Sen (2014) provided a valuable omnibus test for simultaneously checking the assumption of independence between the error and predictor variables and the goodness-of-fit of the parametric model; Rosset and Tibshirani (2018) explored covariate randomness in statistical prediction and applications to covariance penalties; and residual-based goodness-of-fit assessments using a directional test have been explored in Stute (1997).

Nikki L. B. Freeman is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: nlbf@live.unc.edu). Xiaotong Jiang is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: xiaotong@live.unc.edu). Owen E. Leete is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: oleete@email.unc.edu). Daniel J. Luckett is a postdoctoral research associate, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: luckett@live.unc.edu). Teeranan Pokaprakarn is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: terranan@live.unc.edu). Michael R. Kosorok is the W.R. Kenan, Jr. Distinguished Professor and Chair, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA (e-mail: kosorok@bios.unc.edu).

The work of Buja et. al. stands out in its thoroughness of investigation into the interplay between random covariates and regression model misspecification and its proposed paradigm for thinking about regression modeling. Imagining what it would mean to fully adopt the ideas put forth has sparked many lively discussions among us. In our conversations that ranged from the philosophical underpinnings of statistical inference to the practical business of data analysis, we found that Buja et. al., guided us toward important questions but we were unable to fully resolve those questions within their framework. In the following, we detail some of those questions.

2. THE DATA ANALYSIS PIPELINE

Even in our earliest conversations, our attention was drawn to the question of what "Models as Approximations, Parts I and II" means for the real data analysis pipeline. The papers immediately challenge us to critically examine the primary assumptions of statistical modeling and the consequences of when those assumptions are wrong. The authors reference, but do not state, the quote from Box (1979), and we feel it would be instructive to examine the sentiment expressed by Box in greater detail. In his paper, Box disregards the question "Is the model true?" in favor of the question "Is the model illuminating and useful?" This idea was refined to a more practical approach in Box and Draper (1987) where he asks, "How wrong do [models need] to be to not be useful." Much of Part I is dedicated to a rather convincing argument that treating the regressors as fixed can lead to misspecification issues where a model is so wrong that it is no longer useful. While it is true that the ancillarity of the regressor distribution is an assumption frequently made without much justification, the possible negative repercussions are covered in such detail that a cursory reading may leave the reader with a pessimistic view of modeling in general. In many ways, it seems as if the authors focus too much on how modeling needs to change to accommodate potential misspecification rather than identifying the underlying problems and seeking ways to improve the utility of the models we use. In this, we prefer the view

of McCullagh and Nelder (1983), that even though all models are wrong, "some ... are better than others and we can search for the better ones."

We believe that the reweighting technique, presented in Section 5 of Part II, used to diagnose model misspecification by perturbing only the regressor distributions, may provide a useful first step in improving the utility of the models we use. The authors provide a compelling argument for this powerful, intuitive tool for ruling out misspecified candidate models and detecting the presence of interactions without the need to design and test specific models. We think it is worth exploring whether the proposed diagnostics have the potential to improve current model selection procedures, especially for the second-order terms. As the paper pointed out, such diagnostics are not stand-alone analyses or meant to replace existing model diagnosis methods. Rather, they provide additional information about the regressors and model fitting. This leads to the following questions: (1) Under the new view of model specification that the authors put forward, would the current model selection methods (e.g., likelihood ratio test and nested models) still be valid if estimators are not model-robust and might be highly dependent on correct model specification?; (2) Can we combine the traditional model selection methods with the proposed diagnostics? For example, we first apply the diagnostics to detect potential interactions or nonlinearity effects to design candidate models, then test and compare the goodness-of-fit with traditional methods; (3) Will the proposed notion of well-specification and model diagnosis provide a natural ordering of candidate models (such as the traditional step-up and step-down selection procedures) with well-defined measures?

Once a final model is selected, the results need to be conveyed to the investigators, but it is unclear how to interpret a misspecified model. As discussed in Part I, Section 10, not every model will be well-specified, either because misspecification is not detected or because it is being tolerated for insightful simplification. The authors outline a possible method of interpreting a parameter/functional when a nonlinear effect is modeled as linear. Ultimately, this proposed interpretation reduces to whether the general trend is expected to be positive, negative or null. Since nonlinearity could be lurking undetected in any model, it could be argued that all parameters/functionals should be interpreted in this fashion, but we feel that that approach is too conservative. Therefore, let us focus on the scenario where a known nonlinear effect is ignored for the sake of simplicity. The proposed interpretation lacks any information about the magnitude of the effect, but there are several situations where the magnitude represents important information. For example, if the purpose of the model is to determine if intervening on a specific variable will produce a desired result, then knowing only the sign of the effect may be enough; however, the magnitude is useful for determining if the cost of intervening is justified by the expected benefit. It is common to examine several regressors simultaneously with the goal of ranking the relative effects of the regressors. Under the proposed framework the functional could be interpreted as the average effect over the observed regressor distribution, which may still be a useful measure in this context.

3. SEMIPARAMETRIC MODELS

The authors work out the implications of model misspecification, specifically nonlinearity of the conditional mean, for OLS linear regression, and argue for treating the regressors as random as opposed to fixed. Furthermore, the authors argue for a reinterpretation of model parameters as statistical functionals, which depend on the distribution of the regressor distribution. In order to make inference as light on model assumptions as possible, model-robust inference should therefore be used rather than model-trusting inference. Related to this is the idea of semiparametric inference. A semiparametric model has both parametric and nonparametric components, allowing the analyst to make fewer modeling assumptions by focusing on the part of the model that is of interest and leaving other components unspecified. Much of the theory of semiparametric inference can be thought of as seeking the most efficient estimator for the parametric component in the presence of the nonparametric component. This is a very different view of inference with minimal modeling assumptions than the one the authors propose.

In Part II, the authors develop their framework for general parametric regression models, however, semiparametric regression models such as the Cox model are commonly used in practice. A natural extension, then would be to explore the idea of randomness in the regressors and model misspecification in semiparametric regression models and its effect on inference in models with a nonparametric component. The linearity (or proportional hazards) assumption in the Cox model allows for estimation based on the partial likelihood. It is not clear how violation of the proportional hazards assumption would interact with randomness of the regressors in the presence of the nonparametric baseline hazard function. Nonetheless, a number of authors

have studied misspecification in the Cox model. Lin and Wei (1989) proposed a robust procedure similar to those used in the parametric setting and showed that a more complicated form of the "sandwich" variance estimator is valid. Moreover, even if the proportional hazards assumption is violated or if a covariate is omitted, hypothesis tests for β will still be valid. Hence, it seems that, at least in this particular semiparametric model, valid inference is possible under model misspecification such as nonlinearity.

The proportional hazards frailty regression model (Kosorok, Lee and Fine, 2004) is another example of a semiparametric model proposed as an approximation to the truth. Model misspecification and omitted covariates are accounted for by including a frailty term with unknown variance, W, in the model

$$\lambda\{t; Z(t), W\} = a(t)\exp\{\log(W) + \beta' Z(t)\}, \tag{1}$$

where λ is the hazard function, $a(t)$ is the baseline hazard function, and $Z(t)$ is a (possibly time-dependent) covariate. They also show that the direction of β can be correctly estimated even if (1) the frailty distribution scale-family is misspecified or (2) a Cox model is used for estimation but the true model is (1) with a nonzero variance W. This work introduced asymptotic theory for misspecified nonparametric maximum likelihood estimation in semiparametric survival models; the nonparametric component makes this distinct from White's work (1982).

We note that both the examples discussed above treat the regressors as fixed. Therefore, the work that has been done in model-robust inference for the Cox model does not address some of the issues the authors put forth in this paper, such as the need to view the regressors as random. An avenue for future work would be to examine whether the regression parameter, β, in the Cox model can/should be viewed as a statistical functional and whether the ideas of this paper can be extended despite the nonparametric baseline hazard function. It is known that the Cox model can be expressed as a linear transformation model, $h(t) = -\beta' Z + \epsilon$, where t is the failure time, h is some unknown transformation, and ϵ is a random error (Cheng, Wei and Ying, 1995). It would be interesting to examine whether a linear transformation model could be useful in extending the model-robust framework introduced in this paper to semiparametric models. This would allow for developing a ratio of asymptotic variances test statistic for the Cox model and finding new interpretations for the parameters in the Cox model that are valid despite randomness in the regressors.

4. CONCLUSION

In their stimulating work, Buja and his coauthors address the essence of regression modeling. It is kindling, the start to a useful way for thinking about and doing statistical modeling. We have shared in fuller detail here a few of our questions that their work has sparked and that we believe are worthwhile directions for further exploration, but there are others. For example, what are the implications of randomness in the regressors and misspecification in high dimensional data analysis and for post-selection inference? And, what are the implications in a designed study? The implications and what to do about them are not immediately clear in these cases. We hope that the questions we raise are useful for further extensions of this work and for thinking about the problems articulated in these papers. Finally, we again commend the authors for their stimulating work.

REFERENCES

BOX, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.) 201–236. Academic Press, New York.

BOX, G. E. P. and DRAPER, N. R. (1987). *Empirical Model-Building and Response Surfaces*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0861118

CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. MR1380818

KOSOROK, M. R., LEE, B. L. and FINE, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32** 1448–1491. MR2089130

LIN, D. Y. and WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* **84** 1074–1078. MR1134495

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0727836

ROSSET, S. and TIBSHIRANI, R. J. (2018). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *J. Amer. Statist. Assoc.* 1–14.

SEN, A. and SEN, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika* **101** 927–942. MR3286926

STUTE, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25** 613–641. MR1439316

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163