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).

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

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).