Statistical Science
2019, Vol. 34, No. 4, 591–598
https://doi.org/10.1214/19-STS747
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Comment: Models as (Deliberate) Approximations

David Whitney, Ali Shojaie and Marco Carone

1. OVERVIEW

We applaud Buja and coauthors for drawing further attention to the important problem of model misspecification in regression and to the study of its ramifications. In their interesting piece, they advocate for viewing model-based regression coefficients as nonparametric functionals of the data-generating mechanism. This viewpoint has the advantage of clarifying the definition of the estimand and formalizing how to perform model-robust inference based upon influence functions. In our note, we would like to continue the conversation along these lines. We wish to highlight additional considerations that arise in the context of model misspecification in a broader range of scenarios. Our main points are as follows:

(i) the model-robust interpretation of model-based estimands may not always be appealing, particularly when there is significant model misspecification or the sampling scheme includes some form of coarsening;
(ii) when the model fitting procedure involves data-adaptive estimation of nuisances, valid model-robust inference may be much more difficult to achieve;
(iii) these difficulties can be preempted by defining deliberate projection parameters and using suitable non or semiparametric techniques for inference.

2. MODEL-ROBUST INTERPRETATION

Framing regression coefficients as indices for the ‘projection’ of the true regression function onto the specified model is intuitively appealing. In our experience, most practitioners are aware that this is implicitly what they are doing when fitting regression models. However, it must be stressed that not all projections are useful projections. Below, we highlight that model-based regression coefficients may have a poor interpretation when (a) the model used is overly parsimonious, or (b) when the data are subject to some form of coarsening.

2.1 Targeted Versus Indiscriminate Parsimony

A primary reason for the popularity of regression models is their ability to summarize parsimoniously key relationships. However, parsimony can have several impacts on the interpretation of regression coefficients. For example, it can mask effect modification—this occurs if the portion of the model pertaining to the exposure of interest is parsimonious. This may be desirable if the goal is to succinctly summarize population-averaged relationships. This targeted form of parsimony is what renders regression models attractive. However, parsimony could also result in poor confounding control—this occurs when the portion of the model that involves potential confounders is too inflexible to allow sufficient deconfounding. This is an example of indiscriminate parsimony, which is both unnecessary—it can often be mitigated by the use of regression models with parsimonious exposure involvement but flexible confounding adjustment—and possibly harmful.

As an illustration, we expand upon a simple example stemming from the discussion of Section 10 in Part I. There, the authors note that when the underlying associations exhibit symmetry, there may be little to no linear trend. To be concrete, suppose that the data unit consists of the triple (W, X, Y), including a continuous outcome Y, exposure of interest X, and confounder W, generated from data-generating distribution P. Ordinary least-squares (OLS) regression may often be used in this context, with exposure and confounder both included as main terms, and reported upon with the appropriate caveat that the model coefficients represent

David Whitney is Teaching Fellow in Statistics, Department of Mathematics, Imperial College London, Huxley Building, South Kensington Campus, Imperial College, London SW7 2AZ, United Kingdom (e-mail: d.whitney@imperial.ac.uk). Ali Shojaie is Associate Professor, Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, USA (e-mail: ashojaie@uw.edu). Marco Carone is Assistant Professor and Norman Breslow Endowed Faculty Fellow, Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, USA (e-mail: mcarone@uw.edu).