Statistical Science
2020, Vol. 35, No. 3, 391–393
https://doi.org/10.1214/20-STS777
Main article: https://doi.org/10.1214/20-STS720

Comment: Clarifying Endogeneous Data Structures and Consequent Modelling Choices Using Causal Graphs

Erica E. M. Moodie and David A. Stephens

We read with great interest the article by Qian, Klasnja and Murphy (2020), and commend the authors for focusing on principled estimation and providing a quantitative approach to healthcare delivery through mobile devices. The quantitative analyses studied here could have wide-ranging applications that may serve to increase patient empowerment by taking medical monitoring and even intervention out of the clinic and into the home.

Here, we wish to delve into two complementary aspects of the work: first, we attempt to give clarifications concerning the parameter(s) of interest, and second, we provide visualizations of potential scenarios that may help to clarify estimands and when biases due to endogeneity may arise.

1. TREATMENT EFFECTS: ONE, TWO OR TOO MANY?

The authors focus on the setting of (micro)randomized trials, where the treatment of interest is assigned entirely at random. In this setting, one would often expect to be able to perform causal inference, since a key barrier to doing so— confounding—is eliminated thanks to the randomized nature of the treatment assignment. In the motivating HeartSteps study, for instance, the question of interest is to determine the optimal treatment strategy to remind an individual to exercise or not based on their location and recent step activity, with the goal of maximizing steps taken over the next 30 minutes. This question suggests a causal estimand, targeting the effect of the reminder and any modification of the reminder effect by individual covariates.

We attempt here to provide a more precise focus on the estimand in plausible scenarios of interest: a single treatment effect (a ‘main effect model’), an effect that is modified by covariates (an interaction model), or a truly individualized treatment effect characterized by a random slopes model. In such cases, the effect of covariates beyond their modification of treatment are not of primary interest. Looking to equation (15) and Table 2 of Qian, Klasnja and Murphy (2020), the βs are the only parameters of interest, while the random effects b_{i1} will, themselves, also be essential to tailoring treatment recommendations. As we will demonstrate in the next section, sharpening attention to the parameters of interest allows the analyst to step back from the complexities of *all* dependencies within the longitudinal data generating structure, and take note of those most relevant to the scientific question.

Suppose that there does exist heterogeneity in the treatment effect that cannot only be explained by covariates, but rather requires a random slope term. This would imply a treatment strategy that requires knowing or inferring an individual’s random effect prior to being able to implement the treatment strategy. In the setting described in the study, where there are over 150 measures available on average for the participants, this is feasible; however, the strategy would not immediately generalize to future users. Rather, a potentially significant volume of data would first need to be collected to estimate each user’s random slope.

2. VISUALIZATION WITH ACYCLIC GRAPHS: UNDERSTANDING ENDOGENIETY AND CONSEQUENT MODELLING CHOICES

Qian, Klasnja and Murphy (2020) raise a number of interesting scenarios where bias arises even in seemingly simple situations, such as when treatment is randomized. Some of the scenarios raised may be familiar to those with a causal inference background, whereas there are others that are less obvious and perhaps made somewhat less clear without the explicit specification of the estimand of interest. Here, we attempt to clarify, through the visual means provided by causal diagrams (Greenland, Pearl and Robins, 1999), the estimand(s) of interest and possible sources of bias. To emphasize the estimand-focused framework of a causal paradigm, the effects of interest are shown as black arrows, with other conditional dependencies in the data-generative model are shown in grey.

Consider Figure 1, panel A, which follows the notation of Qian, Klasnja and Murphy (2020): a longitudinal setting where X is endogenous. The relationship of interest

Erica E. M. Moodie is Associate Professor, Biostatistics, McGill University, 1020 Pine Ave W., Montreal, Quebec, Canada H3A 1A2 (e-mail: erica.moodie@mcgill.ca). David A. Stephens is Professor, Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Ave W., Montreal, Quebec, Canada H3A 0B9 (e-mail: david.stephens@mcgill.ca).