Perspective
Volume 12:10, 2021

Journal of Biometrics and Biostatistics

ISSN: 2155-6180 Open Access

Estimation of Causal Effects Using Machine Learning

Vivian Diaz*

Department of Biostatistics, University of Weill Cornell Medicine, New York, USA

Editorial Note

In recent decades, the fields of statistical and machine learning have seen a revolution in the development of data-adaptive regression methods that have optimal performance under flexible, sometimes minimal, assumptions on the true regression functions. These developments have impacted all areas of applied and theoretical statistics and have allowed data analysts to avoid the biases incurred under the pervasive practice of parametric model misspecification. In this commentary, I discuss issues around the use of data-adaptive regression in estimation of causal inference parameters. To ground ideas, I focus on two estimation approaches with roots in semi-parametric estimation theory: targeted minimum loss-based estimation (TMLE; van der Laan and Rubin, 2006) and double/debiased machine learning (DML; Chernozhukov and others, 2018). This commentary is not comprehensive, the literature on these topics is rich, and there are many subtleties and developments which I do not address. These two frameworks represent only a small fraction of an increasingly large number of methods for causal inference using machine learning. To my knowledge, they are the only methods grounded in statistical semi-parametric theory that also allow unrestricted use of data-adaptive regression techniques.

A foundational requirement in statistical science is the ability to quantify the uncertainty associated with an estimate. This is one of the reasons why parametric models are popular, even though they are widely accepted to produce biased estimates. When using parametric methods, classical tools such as the central limit theorem and the law of large numbers may be used to obtain convenient asymptotic approximations to the sampling distribution of the estimators, readily yielding formulas for computing uncertainty measures such as confidence intervals. Despite this convenience, the formulas derived are defective under model misspecification: they correctly quantify the uncertainty around an incorrect target value.

One of the pillars of statistical science, which sets it apart from other data analysis fields, is the ability to perform formal statistical inference. TMLE and DML are examples of methods that successfully bridge the gap between machine learning and statistical science, allowing data analysts to use machine learning while obtaining valid methods for formal statistical inference. TMLE and DML achieve this by embracing the technical challenges that come with realistic statistical models, instead of shunning them in favor of the analytical convenience of unscientific parametric methods. In a big data world dominated by complex causal inference questions and machine learning applications, the continued success of statistics as a profession depends on its ability to adapt to these new tools and respond to real-life scientific problems while being truthful to its foundational principles. Formulating and answering meaningful scientific problems requires causal inference thinking. In-corporating machine learning into our estimation toolbox while performing statistical inference requires the integration of empirical processes, high-dimensional statistics, and semi-parametric and non-parametric statistics. I believe these points should be kept in mind as our statistics discipline goes through the changes brought about by the advent of data science, particularly as these changes relate to PhD and MS curricula and data analysis quality standards for applied research.

How to cite this article: Diaz, Vivian. "Estimation of Casual Effects Using Machine Learning." *J Biom Biostat* 12 (2021) : 523.

***Corresponding Author:** Vivian Diaz, Department of Biostatistics, University of Weill Cornell Medicine, New York, USA, Tel: +18526457820; E-mail: viviandiaz@gmail.com*

Received: October 05, 2021; **Accepted**: October 19, 2021; **Published**: October 26, 2021