The Journal of Open Source Software

causal-curve: A Python Causal Inference Package to Estimate Causal Dose-Response Curves

Roni W. Kobrosly[1, 2]

1 Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA **2** Flowcast, 44 Tehama St, San Francisco, CA, USA

DOI: 10.21105/joss.02523

Software

- Review
- Repository
- Archive

Editor: Olivia Guest

Reviewers:

- @cmparlettpelleriti
- @tomfaulkenberry
- @alexjonesphd

Submitted: 08 July 2020
Published: 27 July 2020

Summary

In academia and industry, randomized controlled experiments (colloquially "A/B tests") are considered the gold standard approach for assessing the impact of a treatment or intervention. However, for ethical or financial reasons, these experiments may not always be feasible to carry out. "Causal inference" methods are a set of approaches that attempt to estimate causal effects from observational rather than experimental data, correcting for the biases that are inherent to analyzing observational data (e.g. confounding and selection bias) (Hernán & Robins, 2020).

Although significant research and implementation effort has gone towards methods in causal inference to estimate the effects of binary treatments (e.g. what was the effect of treatment "A" or "B"?), much less has gone towards estimating the effects of continuous treatments. This is unfortunate because there are a great number of inquiries in research and industry that could benefit from tools to estimate the effect of continuous treatments, such as estimating how:

- the number of minutes per week of aerobic exercise causes positive health outcomes, after controlling for confounding effects.
- increasing or decreasing the price of a product would impact demand (price elasticity).
- changing neighborhood income inequality (as measured by the continuous Gini index) might or might not be causally related to the neighborhood crime rate.
- blood lead levels are causally related to neurodevelopment delays in children.

`causal-curve` is a Python package created to address this gap; it is designed to perform causal inference when the treatment of interest is continuous in nature. From the observational data that is provided by the user, it estimates the "causal dose-response curve" (or simply the "causal curve").

In the current release of the package there are two unique model classes for constructing the causal dose-response curve: the Generalized Propensity Score (GPS) and the Targetted Maximum Likelihood Estimation (TMLE) tools. There is also tool to assess causal mediation effects in the presence of a continuous mediator and treatment.

`causal-curve` attempts to make the user-experience as painless as possible:

- This package's API was designed to resemble that of `scikit-learn`, as this is a commonly used Python predictive modeling framework familiar to most machine learning practitioners.
- All of the major classes contained in `causal-curve` readily use Pandas DataFrames and Series as inputs, to make this package more easily integrate with the standard Python data analysis tools.

- A full, end-to-end example of applying the package to a causal inference problem (the analysis of health data) is provided. In addition to this, there are shorter tutorials for each of the three major classes are available online in the documentation, along with full documentation of all of their parameters, methods, and attributes.

This package includes a suite of unit and integration tests made using the pytest framework. The repo containing the latest project code is integrated with TravisCI for continuous integration. Code coverage is monitored via codecov and is presently above 90%.

Methods

The `GPS` method was originally described by Hirano (Hirano & Imbens, 2004), and expanded by Moodie (Moodie & Stephen, 2010) and more recently by Galagate (Galagate, 2016). GPS is an extension of the standard propensity tool method and is essentially the treatment assignment density calculated at a particular treatment (and covariate) value. Similar to the standard propensity score approach, the GPS random variable is used to balance covariates. At the core of this tool, generalized linear models are used to estimate the GPS, and generalized additive models are used to estimate the smoothed final causal curve. Compared with this package's TMLE method, this GPS method is more computationally efficient, better suited for large datasets, but produces significantly wider confidence intervals.

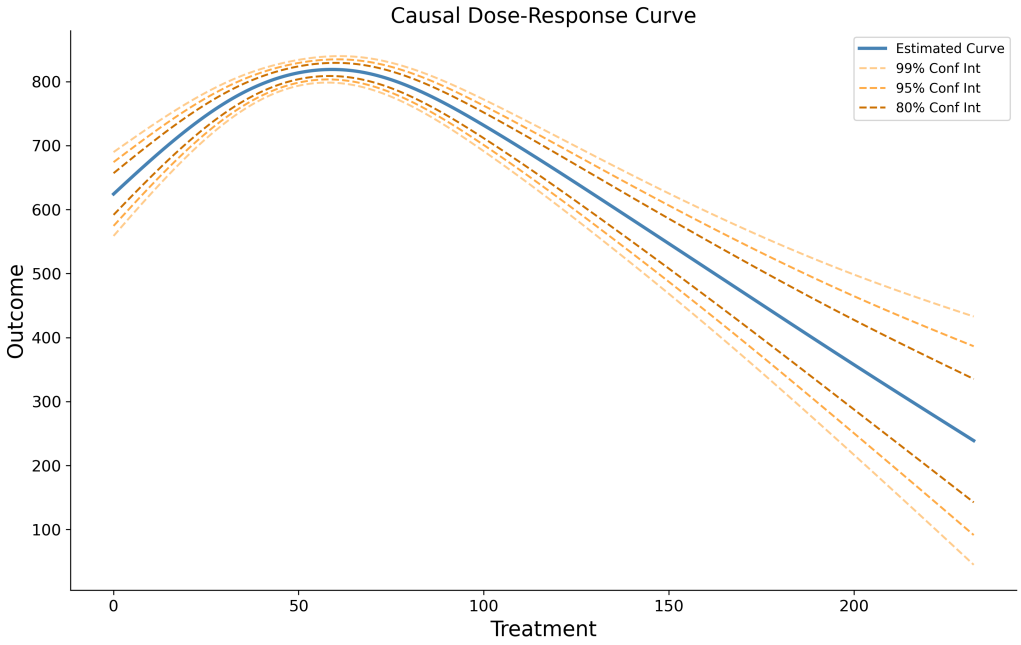

Figure 1: Example of a causal curve generated by the GPS tool.

The `TMLE` method is based on van der Laan's work on an approach to causal inference that would employ powerful machine learning approaches to estimate a causal effect (van der Laan & Gruber, 2010). TMLE involves predicting the outcome from the treatment and covariates using a machine learning model, then predicting treatment assignment from the covariates. TMLE also employs a substitution "targeting" step to correct for covariate imbalance and to estimate an unbiased causal effect. Currently, there is no implementation of TMLE that is suitable for continuous treatments. The implemention in `causal-curve` constructs the final curve through a series of binary treatment comparisons across the user-specified range of

treatment values and then by connecting them. Compared with the package's GPS method, this TMLE method is double robust against model misspecification, incorporates more powerful machine learning techniques internally, produces significantly smaller confidence intervals, however it is less computationally efficient.

`causal-curve` allows for continuous mediation assessment with the `Mediation` tool. As described by Imai this approach provides a general approach to mediation analysis that invokes the potential outcomes / counterfactual framework (Imai & Tingley, 2010). While this approach can handle a continuous mediator and outcome, as put forward by Imai it only allows for a binary treatment. As mentioned above with the `TMLE` approach, the tool creates a series of binary treatment comparisons and connects them to show the user how mediation varies as a function of the treatment. An interpretable, overall mediation proportion is provided as well.

Statement of Need

While there are a few established Python packages related to causal inference, to the best of the author's knowledge, there is no Python package available that can provide support for continuous treatments as `causal-curve` does. Similarly, the author isn't aware of any Python implementation of a causal mediation analysis for continuous treatments and mediators. Finally, the tutorials available in the documentation introduce the concept of continuous treatments and are instructive as to how the results of their analysis should be interpretted.

Acknowledgements

We acknowledge the valuable feedback from Miguel-Angel Luque, Erica Moodie, and Mark van der Laan during the creation of this project.

References

Galagate, D. (2016). *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response function with applications.* Digital Repository at the University of Maryland.

Hernán, M., & Robins, J. (2020). *Causal Inference: What If.* Chapman & Hall.

Hirano, K., & Imbens, G. (2004). *The propensity score with continuous treatments*. Wiley.

Imai, K., K., & Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychological Methods*, *15*. doi:10.1037/a0020761

Moodie, E., & Stephen, D. (2010). Estimation of dose–response functions for longitudinal data using the generalised propensity score. *Statistical Methods in Medical Research*, *21*. doi:10.1177/0962280209340213

van der Laan, M., & Gruber, S. (2010). Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*. doi:10.2202/1557-4679.1181