Machine Learning Methods for Causal Inference with Continuous Treatments

Eray Turkel (SUID: eturkel)

June 11, 2019

We propose and test the accuracy of various machine learning methods for estimating the causal effects of continuous treatments. The literature on treatment effect estimation has been mostly focused on making causal inferences with binary treatments (Athey (2015), Athey and Imbens (2015)): $T_i = 1$ or $T_i = 0$. However, most interesting real life studies involve continuous or multi-valued treatments, where the object of interest is $E(Y(t))$, the expected outcome given the treatment t, which is referred to as the 'dose-response function'. A researcher could be interested in evaluating the effect of different dosages of a drug, or the effect of a training program offered at different lengths (Kluve et al. (2012)), or the causal effect of various levels of ad exposure on purchases or browsing behavior.

In this setting, the model used in the binary treatment framework can be generalized to calculate a 'generalized' propensity score, $P(T = t|X = x)$. This object can be calculated by estimating conditional probability density of the treatment given the covariates, $f_{T|X}(t|x)$. This generalized propensity score can then be used to de-bias and estimate the unbiased population dose response function, $E(Y(t))$, as in the case of binary treatments.

In practice, this conditional distribution is usually estimated by making some parametric and distributional assumptions: a common assumption is the conditional density being a Gaussian distribution: $T_i|X_i \sim N(\beta_0 + \beta_1' X_i, \sigma^2)$, and the parameters estimated through simple maximum likelihood. The outcome model $E(Y_i|T_i, P_i)$ is usually estimated by ordinary least squares, adding high order interactions of the dosage T_i and the calculated generalized propensity score P_i (see Hirano and Imbens (2004), Imbens (2000)).

This procedure consists of two separate 'prediction' steps, and machine learning algorithms are likely to perform better than the older methods used in the literature. The first step is the estimation of the conditional probability density of the treatment given the covariates, $f_{T|X}(t|x)$. For this, we will use conditional kernel density estimators (also called Parzen Window Estimators) for estimating the conditional density of being assigned a dose $T_i = t$ given covariates $X_i = x_i$. The second step is predicting the outcome Y_i given T_i and the probability estimated in the first step. For this step, we will compare the performance of neural networks, gradient boosted tree methods and more standard methods based on linear regression that are common in the literature.

First, we conduct simulation experiments for the second step: predicting the outcome Y_i given T_i and the propensity scores estimated in the first step. For the first step (density estimation), we will use an oracle estimator to get the conditional density of T_i given X_i. This will allow us to compare the performance of various estimators without losing accuracy due to misspecifications in the first step. Moreover, density estimation is a considerably harder task given the high dimensional nature of the problem. In the next section where we apply our method to real data, we will use kernel density estimators.

1 Numerical Experiments

We conduct simulation exercises using a known data generating process to evaluate the accuracy of the proposed methods. We generate data according to the following distribution:

$$X_1, X_2, ..., X_p \sim N(0,1), i.i.d., T_i = Max(\alpha X_1 + \nu, 0), \nu \sim N(0,1)$$

Where T_i the treatment is distributed according to a truncated normal distribution with mean αX_1. We will allow T to depend on more covariates later on. We then generate the outcome Y_i according to the following process:

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \beta_{p+1} X_1^2 + \beta_{p+2} X_1 X_2 ... + \eta T_i + \epsilon, \epsilon \sim N(0,1), \beta_j \sim N(0,1), \forall j.$$

So the true model contains all of the covariates and all two-way interactions between them, plus the treatment. η measures the true effect of the treatment. We pick each of the coefficients β_j for the covariates from a normal distribution. We then set a certain fraction of these coefficients to zero, to test the model with varying degrees of sparsity. Generally, given a fixed level of noise in the underlying process, increasing sparsity improves the performance of simpler models. This could be thought as a way of controlling the noise-to-signal ratio in the simulation study: if X_i's enter the model in a sparse way, it will be easier to distinguish the effect of T_i from the noise. With a very dense model, due to confounding, the effect of T_i can be missed. In this simple setup, the true dose response function is linear in t: $E(Y(t)) = \eta t$.

Since $T_i = Max(\alpha X_1 + \nu, 0)$ is a truncated normal, the propensity becomes:

$p(T_i = t|X) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(t-\alpha X_1)^2}{2\sigma^2})$, if $t > 0$, and

$p(T_i = t|X) = \Phi(-\alpha X_1/\sigma)$, if $t = 0$, where Φ is the CDF of the normal distribution.

We run a truncated linear regression of T on all the covariates X using maximum likelihood, and get an estimate of the coefficient α[1]. Using this coefficient, we calculate

[1]Note that this approach technically allows the conditional mean of the distribution to depend on covariates other than X_1 as well. This will be useful in our numerical experiments where we increase the

$\eta = 4, \alpha = 0.8$	Flexible Linear Regression	Gradient Boosting	Neural Network
Sparsity= 0.05	58.66	55.59	52.31
Sparsity=0.5	10.67	12.33	18.67
Sparsity=0.9	5.17	2.83	3.73

Table 1: Average MSE of the models in 100 simulations in each scenario. Underlying response function is linear in t, with low amount of confounding with a single variable. $E(Y(t)) = 4t$ and $t|X \sim truncNorm(\mu = 0.8X_1)$.

the propensities evaluating the density of the truncated normal, as explained above.

Given the propensity score, we move on to the estimation of $E(Y_i(t))$, the dose response function given the dosage t. Hirano and Imbens (2004) suggest a flexible linear model. Given a very dense model with a large number of covariates, it will be impossible to control for all of the covariates in the outcome model. They show that controlling for the propensity score P_i and T_i is enough to get an unbiased estimator of $E(Y(t))$. First, we model the outcome process as:

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 P_i + \alpha_4 P_i^2 + \alpha_5 P_i T_i$$

and estimating the dose response function using the fitted model at every point t using:

$$\hat{E}(Y(t)) = 1/N \sum_{i=1}^{N} \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 \hat{P}(t, X_i) + \alpha_4 \hat{P}(t, X_i) + \alpha_5 \hat{P}(t, X_i) t$$

Where $P(t, X_i)$ is the PDF evaluated at t and X_i. As a baseline, we implement this strategy that is widely used in the literature. We will compare this output model against Neural Nets and Gradient Boosted Trees.

For the Neural Net, we use a two hidden-layer network with 4 and 2 nodes at each layer, with the TanH activation function and a linear output. We use He initialization and train the model using (unbatched) gradient descent, for 100,000 iterations. We train each model 5 times with different random initializations and take the best output. For the GBM, we use 5000 trees, and use cross-validation to determine the optimal number of trees to use at prediction time on test data. In all three models, we use an unseen test set generated through the same process to evaluate the accuracy of the predictions.

degree of confounding by allowing treatment to increase with other covariates.

Complex Model	Flexible Linear Regression	Gradient Boosting	Neural Network
Sparsity=0.05	32.81	9.88	7.43
Sparsity= 0.5	29.24	7.07	8.63
Sparsity=0.9	35.63	4.29	6.80

Table 2: Average MSE of the models in 100 simulations in each scenario. Underlying response function is non-linear in t, with a complex confounding model. $E(Y(t)) = t + 1/2t^2 + 1/8t^3$, and $t|X \sim truncNorm(\mu = X_1 + 5X_2 + 0.5X_3)$

Overall, we see the flexible linear model to perform well in simple sparse setups, but gradient boosted methods and neural nets perform much better in denser and more complicated models. We plot an example from the nonlinear dense scenario (sparsity=0.05) for demonstrative purposes. The flexible linear regression model with higher order terms is biased and fails to capture the true dose response, while the gradient boosting model performs the best.

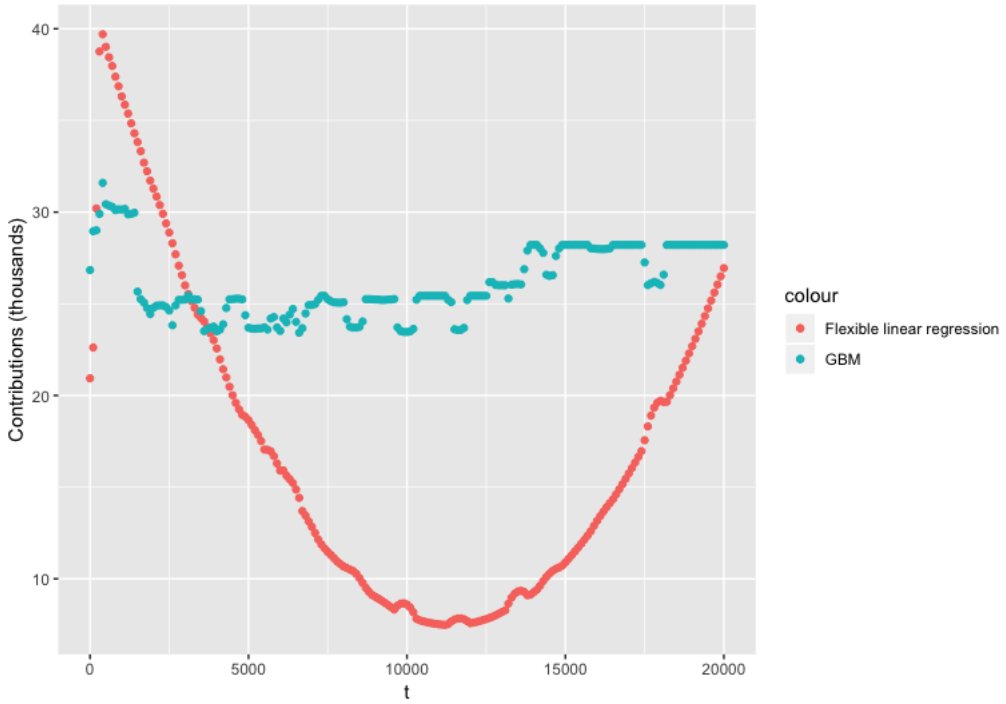
2 Application to Real Data

In this section, we apply our method to a real life study on evaluating the causal effect of political ads on money raised by candidates. The data comes from the replication archive of Urban and Niebler (2014) and contains detailed information on the demographics and the political characteristics of 16,000 US Zipcode areas in 2008. The covariates we include (following the authors) are population shares of black, white, and hispanic groups within each zipcode, share of the population over the age of 65, population share of college graduates, population density, per capita household income, percent of people who voted in the previous election (2004), a dummy variable for Rural zipcodes, and the republican voteshare in the last presidential election (2004).

Political candidates buy TV ads based on designated market areas (DMAs), which are geographical regions that encompass multiple counties and sometimes state borders. Also, because of the electoral college system, presidential candidates focus on competitive 'battleground' states. While some states get treated very intensely by political ads, in other non-battleground states, candidates don't buy a single ad. Borrowing from the authors, as an example, "areas of northeastern New York receive spillover ads from Vermont and New Hampshire, while areas of northern and western Texas receive spillover ads from New Mexico". Because of these mismatches between county borders and DMA boundaries, some states and counties get targeted by ads 'accidentally', because they share a border with a competitive state. The authors argue that these 'accidental' exposures to political ads are as good as exogenous once we condition for the covariates, and we can use these ad exposures to measure the effect of ads on money raised by candidates. The main outcome variable is the total amount of contributions (above 200 dollars) originating from that zipcode.

The authors use the number of ads aired (30-second TV spots) in each zipcode as their

main treatment variable. It is a continuous variable ranging from 0 to 25000, however, they discretize this and analyze it as a binary treatment, with the range (0,1000) as the 'untreated' case, and the range $(1000, \infty)$ as the 'treated' case. This naturally leads to loss of information and the causal interpretation of the resulting estimates is not very clear. The authors report the average treatment effect on the treated as approximately \$6,200. This is the effect of airing more than 1000 ads in a zipcode. The effect is surprisingly big, given that the average contributions from a zipcode is about \$22,000. The authors go on to suggest that presidential candidates are 'leaving dollars on the sidewalk' by not airing ads in noncompetitive states, given that airing ads results in such an increase in contributions.



We use our proposed method to estimate the response function in this context. We calculate the conditional density of the treatment, conditioning on 10 covariates defined above. We split the data to 3 groups. The sizes are 8000,3000, 5000 for training, validation, and test, respectively. We use a conditional kernel density estimator and pick the optimal bandwidths through cross-validated maximum likelihood, using 50 restarts and an Epanechnikov kernel on the training set. For the GBM, we train the model using 15000 trees and use 10-fold cross validation to pick the number of trees to use in the test set at prediction time.

Overall, our results suggest significant returns to advertising in terms of money raised. The GBM approach seems much better suited to the task at hand compared to the standard flexible linear approach. The GBM results are also more interpretable: we see increasing returns to advertising, where airing up to 13,000 ads increases donations by 25,000 USD, and airing more than 13,000 ads increases donations by a slightly larger amount, 28,000 USD.

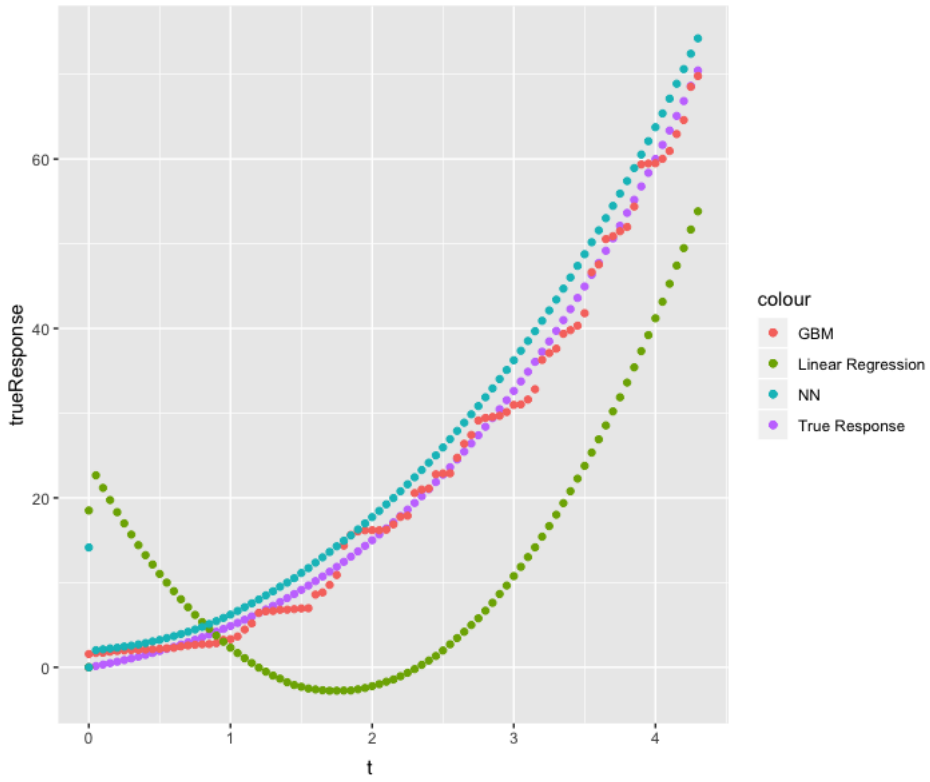
The linear regression model would suggest that airing 10,000 ads is less useful than airing 5000 ads. While this might be justified by decreasing returns to scale, it seems unlikely that candidates would continue to advertise so much if the returns were decreasing. It is clear that the proposed method improves upon the existing linear approach by fitting the data using a more flexible approach. The flexibility allowed by the linear model only allows polynomial terms and interactions, which is not enough to capture the underlying variation in the data. Our results are also significantly different from the results in the paper, where the authors find an average return of 6000-10000 USD of airing ads, using their discretized treatment variable. It is clear that the binary approach leads to loss of information.

3 Code

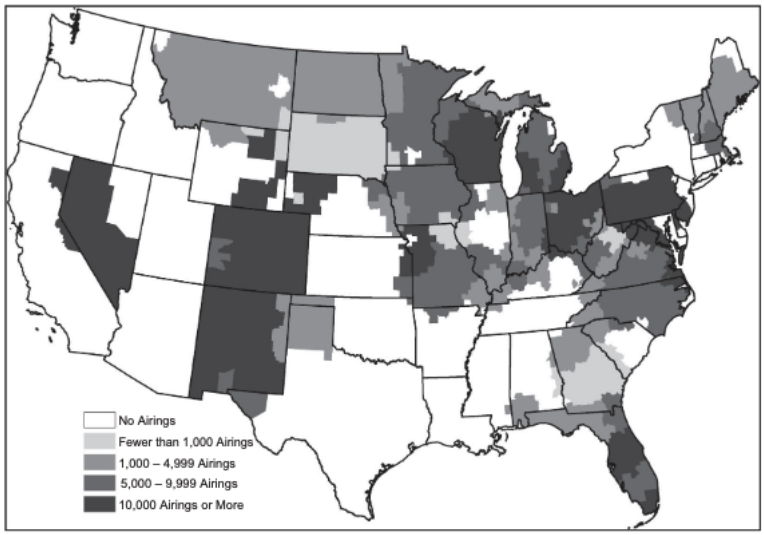
https://github.com/erayturkel/MLContinuous

4 Appendix

Example Simulated Scenario. The true response function is plotted in purple.



Map of the treatment intensities, from Urban and Niedler, 2014. Only non-battleground states that border competitive states and received ads 'accidentally' are included in the data.



References

Athey, S. (2015): "Machine learning and causal inference for policy evaluation," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 5–6.

Athey, S. and G. W. Imbens (2015): "Machine learning methods for estimating heterogeneous causal effects," *Arxiv:stat*, 1050.

Ding, P., F. Li, et al. (2018): "Causal inference: A missing data perspective," *Statistical Science*, 33, 214–237.

Hirano, K. and G. W. Imbens (2004): "The propensity score with continuous treatments," *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164, 73–84.

Holland, P. W. (1986): "Statistics and causal inference," *Journal of the American statistical Association*, 81, 945–960.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012): "Evaluating continuous training programmes by using the generalized propensity score," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 587–617.

Lee, B. K., J. Lessler, and E. A. Stuart (2010): "Improving propensity score weighting using machine learning," *Statistics in medicine*, 29, 337–346.

Rubin, D. B. (2005): "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100, 322–331.

Urban, C. and S. Niebler (2014): "Dollars on the Sidewalk: Should US Presidential Candidates Advertise in Uncontested States?" *American Journal of Political Science*, 58, 322–336.