Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017

P. Richard Hahn[*], Vincent Dorie, and Jared S. Murray

June 4, 2018

Abstract

This brief note documents the data generating processes used in the 2017 Data Analysis Challenge associated with the Atlantic Coast Causal Inference conference. The focus of the challenge was estimation and inference for conditional average treatment effects (CATEs) in the presence of targeted selection, which leads to strong confounding.

1 Introduction

The purpose of the Data Analysis Challenge is to provided a blinded empirical evaluation of causal inference methods on synthetic data, where the true data generating process is revealed after estimation and inference accuracy for each method has been determined. While any given simulation study is necessarily limited, the separation between the data generators and the method submitters here lends additional credibility to the results. The options for empirically evaluating causal methods (generating counterfactual predictions) are also more limited than for traditional (factual) prediction problems, where performance can be evaluated on test subsets of "real" data. The goal here is to design high-quality simulation studies, in the sense that the synthetic data are plausibly representative of real data in key respects.

2 Overview

For the 2017 Challenge, the data were generated according to 32 distinct, fixed, data generating processes (DGPs). For each of these, 250 independent replicate data sets were produced,

[*] Thanks to the Booth School of Business for financial support. Thanks also to Jinhui Xu, Yao Shi, and Changhan He for typesetting assistance.

for a total of 8,000 data sets. All 32 DGPs had covariate-dependent treatment effects. We assumed that

$$\mathrm{E}(Y_i \mid X_i = \mathrm{x}_i, Z_i = z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i) z_i. \tag{1}$$

From a structural equation vantage point, our response variable arises as

$$Y_i = F(\mathrm{x}_i, z_i, \epsilon_i),$$

with treatment Z_i arising as

$$Z_i = G(\mathrm{x}_i, \nu_i)$$

for fixed, known functions F and G. Exogenous noise variables were generated $\nu_i \perp\!\!\!\perp \epsilon_i \perp\!\!\!\perp Z_i$. In terms of potential outcomes we assume that

$$\mathrm{E}(Y^1 \mid \mathrm{x}) = \mu(\mathrm{x}) + \tau(\mathrm{x}), \tag{2}$$

and

$$\mathrm{E}(Y^0 \mid \mathrm{x}) = \mu(\mathrm{x}), \tag{3}$$

so that

$$\mathrm{E}(Y^1 - Y^0 \mid \mathrm{x}) = \tau(\mathrm{x}) \tag{4}$$

where Y^1, Y^0 are the potential outcomes under treatment and control, respectively, so that $\tau(\mathrm{x})$ explicitly represents the CATE. Strong ignorability holds in every scenario considered here. Further, the stronger condition of no unmeasured treatment moderation holds – that is, there are no unmeasured factors that drive treatment effect heterogeneity.

The 32 distinct DGPs were arrived at by considering four different types of errors:

- additive, independent, identically distributed,
- additive, group correlated,
- additive, heteroskedastic,
- non-additive, independent, identically distributed.

The error terms were generated according to Gaussian distributions in all cases.

Within each of these four error types, we considered "high" and "low" settings of three separate aspects of the DGP:

- magnitude of the causal effect,
- strength of the confounding,
- noise level in the response variable.

Interesting aspects of causal DGPs that were not considered this year include:

- non-Gaussian errors,
- high dimensional covariates,
- null effects.

3 Control and/or moderating variables

Across all 8,000 data sets, the measured control and moderating variables remain fixed, taken from the Infant Health and Development Program, or IHDP [Brooks-Gunn et al., 1992]. These covariates were also used in the 2016 ACIC Data Analysis Challenge [Dorie et al., 2017]. Only eight covariates out of 58 from original data were used:

- X_1: Mother's age (continuous),
- X_3: Mother's cigarettes per day (continuous),
- X_{10}: Mother's endocrine condition (binary),
- X_{14}: Mother's nervous system condition (binary),
- X_{15}: Mother's obstetric complications (binary),
- X_{21}: Mother's birth place (categorical),
- X_{24}: Mother's race (binary),
- X_{43}: Child's bilirubin (continuous).

The correlations between these variables are shown in Table 1 for reference.

correlation	X_1	X_3	X_{10}	X_{14}	X_{15}	X_{21}	X_{24}	X_{43}
X_1	1.00	0.04	-0.07	-0.03	-0.04	-0.07	0.03	-0.01
X_3	0.04	1.00	-0.02	0.03	-0.02	-0.10	-0.16	0.13
X_{10}	-0.07	-0.02	1.00	0.04	0.09	-0.02	-0.10	-0.07
X_{14}	-0.03	0.03	0.04	1.00	0.09	-0.03	-0.08	0.07
X_{15}	-0.04	-0.02	0.09	0.09	1.00	-0.03	0.04	-0.04
X_{21}	-0.07	-0.10	-0.02	-0.03	-0.03	1.00	0.20	-0.00
X_{24}	0.03	-0.16	-0.10	-0.08	0.04	0.20	1.00	-0.11
X_{43}	-0.01	0.13	-0.07	0.07	-0.04	-0.00	-0.11	1.00

Table 1: Correlation matrix of control variables.

4 Data generation details

4.1 Additive errors

We denote the varying settings by variables ξ, η, and κ. The variable ξ modulates the magnitude of the effect size, and takes one of two values, 2 or 1/3. The variable η modulates the standard deviation of the error, and takes one of two values, 5/4 or 1/4. The vector variable κ corresponds to regression coefficients governing the propensity to receive treatment, and takes values (3, -1) or (0.5, 0). According to the values of (ξ, η, κ), one obtains eight cases, as shown in Table 2.

No.	Effect magnitude ξ	Noise level η	Selection strength (κ_1, κ_2)
1.	Low (1/3)	Low (0.25)	Weak (0.5 , 0)
2.	Low (1/3)	Low (0.25)	Strong (3 , -1)
3.	Low (1/3)	High (1.25)	Weak (0.5 , 0)
4.	Low (1/3)	High (1.25)	Strong (3 , -1)
5.	High (2)	Low (0.25)	Weak (0.5 , 0)
6.	High (2)	Low (0.25)	Strong (3 , -1)
7.	High (2)	High (1.25)	Weak (0.5 , 0)
8.	High (2)	High (1.25)	Strong (3 , -1)

Table 2: Eight different cases and the values of corresponding parameters.

Next, define the following functions:

$$\begin{aligned} f(\mathrm{x}) &= x_1 + x_{43} + 0.3(x_{10} - 1), \\ \pi(\mathrm{x}) &= \Pr(Z_i = 1) = (1 + \exp(\kappa_1 f(\mathrm{x}) + \kappa_2))^{-1}, \\ \mu(\mathrm{x}) &= -\sin(\Phi(\pi(\mathrm{x}))) + x_{43}, \\ \tau(\mathrm{x}) &= \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1)), \\ \sigma(\mathrm{x}) &= 0.4 + \frac{x_{21} - 1}{15}, \end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. Let $\sigma_y = \eta\sqrt{\mathrm{Var}(\mu(\mathrm{x}) + \pi(\mathrm{x})\tau(\mathrm{x}))}$, where the variance is taken over the observed sample. Finally, let $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, 1)$. With these definitions in hand, we can now describe the data generation protocols as follows.

- In the *independent, identically distributed errors* case we simply have

  $$Y_i = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i)Z_i + \sigma_y \varepsilon_i.$$

- In the *group correlated errors* case we have

  $$Y_i = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i)Z_i + \sigma_y(0.9\varepsilon_i + 0.1\epsilon_{x_{i,21}}),$$

  where $\epsilon_{x_{i,21}}$ denotes one of sixteen realizations of a standard normal random variable, indexed by the variable x_{21}, which takes sixteen levels (recorded as 1 through 16). Accordingly, 10% of the error term is shared among variables with common values of x_{21} (mother's place of birth).

- In the *heteroskedastic error* case, we have

  $$Y_i = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i)Z_i + \sigma(\mathrm{x}_i)\sigma_y \varepsilon_i,$$

  so that the error variance varies as a function of x_{21} (mother's place of birth). Observe that $\sigma(\mathrm{x})$ ranges from 0.4 to 1.4 in constant increments.

4.2 Non-additive errors

The notation for the non-additive case is considerably more cumbersome. Operationally, we simply generate data according to an additive error model and then perform a nonlinear transformation. However, in doing so we must correct for the causal effect, which is no longer explicitly defined, as in the additive case. Secondly, we endeavored to select the numerical constants governing the transformation so that the distribution of the outcome variable did not look dramatically different than it did in the additive cases. That is, the data is generated according to

$$\begin{aligned}\tilde{Y}_i &= \tilde{\mu}(\mathrm{x}_i) + \tilde{\tau}(\mathrm{x}_i)Z_i + \sigma_y\varepsilon_i,\\ Y_i &= 13\Phi(\tilde{Y}_i \mid a, b) - 6,\end{aligned}$$

where $\Phi(\cdot \mid a, b^2)$ denotes the CDF of a $\mathrm{N}(a, b^2)$ random variable and the functions are defined as above. Note that for $\tilde{\mu}(\mathrm{x})$ and $\tilde{\tau}(\mathrm{x})$ we use the definitions given above for $\mu(\mathrm{x})$ and $\tau(\mathrm{x})$; the new notation highlights that these functions no longer bear a direct relationship to the potential outcomes as described in Section 2. Next, we determine specific values of a and b as:

$$\begin{aligned}a &= \mathrm{E}(\tilde{Y}) = \frac{1}{n}\sum_i \tilde{\mu}(\mathrm{x}) + \tilde{\tau}(\mathrm{x})\pi(\mathrm{x})\\ b &= 1.25\sqrt{\mathrm{Var}(\tilde{Y})} = 1.25\sqrt{\sigma_y^2 + \mathrm{Var}(\tilde{\mu}(\mathrm{x}) + \tilde{\tau}(\mathrm{x})\pi(\mathrm{x})}\end{aligned}$$

where the variance is taken over the sample.

Finally, to compute the CATE, we employ the following fact concerning a standard normal random variable $W \sim \mathrm{N}(0, 1)$:

$$\mathrm{E}(\Phi(m + sW)) = \Phi\left(\frac{m}{\sqrt{1+s^2}}\right).$$

Applying this to our specific case, we write

$$\begin{aligned}m^1(\mathrm{x}_i) &= \frac{\tilde{\mu}(\mathrm{x}_i) + \tilde{\tau}(\mathrm{x}_i) - a}{b},\\ m^0(\mathrm{x}_i) &= \frac{\tilde{\mu}(\mathrm{x}_i) - a}{b},\\ s &= \sigma_y/b,\end{aligned}$$

and find that

$$\begin{aligned} \mathrm{E}(Y^1 \mid \mathrm{x}) &= 13\Phi\left(\frac{m^1(\mathrm{x})}{\sqrt{\frac{\sigma_y^2}{b^2}+1}}\right) - 6, \\ \mathrm{E}(Y^0 \mid \mathrm{x}) &= \mu(\mathrm{x}) \;=\; 13\Phi\left(\frac{m^0(\mathrm{x})}{\sqrt{\frac{\sigma_y^2}{b^2}+1}}\right) - 6. \end{aligned}$$

It follows that the treatment effect can be calculated as

$$\tau(\mathrm{x}) = \mathrm{E}(Y^1 \mid \mathrm{x}) - E(Y^0 \mid \mathrm{x}) = 13\Phi\left(\frac{m^1(\mathrm{x})}{\sqrt{\frac{\sigma_y^2}{b^2}+1}}\right) - 13\Phi\left(\frac{m^0(\mathrm{x}_i)}{\sqrt{\frac{\sigma_y^2}{b^2}+1}}\right).$$

5 Targeted selection

The main theme (unannounced) of the challenge this year is the notion of "targeted selection". That is, we were interested in understanding the behavior of various methods in situations where the likelihood that an individual receives treatment is a function of the expected response of that individual if left untreated. In the DGPs above this manifest as $\mu(\mathrm{x})$ being a function of $\pi(\mathrm{x})$; in retrospect it would have been more clearly experessed the other way around, with $\pi(\mathrm{x})$ being written as a noisy function of $\mu(\mathrm{x}) = \mathrm{E}(Y^0 \mid \mathrm{x})$. This type of confounding structure is highly plausible in many real world scenarios where the course of treatment is directly predicated on a prognosis determined by observed covariates, yet it is unlikely to arise from DGPs that are generated stochastically using polynomial bases (such as the 2016 Data Challenge). We hope that others are encouraged to investigate this type of DGP in future methodological development.

6 Data files

6.1 File structure

The exact data files used are available in the contest˙data folder, stored with the following file structure. There are four folders, named

- `group_corr`

- heteroskedastic
- iid
- non-additive.

Within each folder are eight subfolders, named according to a binary encoding of the simulation settings — the first bit corresponds to the effect magnitude parameter ξ, the second bit corresponds to the noise level parameter η, and the third bit corresponds to the selection strength parameter vector κ. In all cases, a zero corresponds to the low setting and a 1 corresponds to the high setting.

Within each of these subfolders there are 253 individual comma separated value files (.csv). The files named 1.csv through 250.csv correspond to the 250 replicated data sets of the given DGP. The files named 1.***.y.csv and 1.***.z.csv are demo files that were provided to participants prior to submission; the response variable file was generated using a *different* set of treatment variables than are given in the treatment indicator file. Finally, the dgp.csv file contains two columns, one of which contains the (conditional) treatment effect (somewhat confusingly referred to as "alpha" in the files, whereas "tau" would make better sense relative to this write-up) and the other which contains the corresponding mean of the untreated potential outcome (μ(x)).

NOTE: There is an error in the data files generated with heteroskedastic errors. The treatment effect for these files was, erroneously, calculated according to the formulas for the non-additive error case. We are leaving these files for documentation purposes, but they are incorrect. The results for the heteroskedastic errors are likewise incorrect and we omit them from further consideration.

6.2 Covariate matrices

Common to all of these files is a shared covariate matrix, given as X.csv. It contains $n = 4,302$ rows, corresponding to individuals in the original IHDP data set. This number is smaller by 500 than the covariate matrix used in the 2016 Data Analysis Challenge; 250 observations were provided to participants in advance and were used to generate the response variables given in the 1.***.y.csv files and another 250 were used similarly to generate the 1.***.z.csv files. These matrices are called X_subset_y.csv and X_subset_z.csv respectively. Additionally, the data generating processes described above actually utilized a transformed set of covariate values (the same as were used in the 2016 Data Analysis Challenge); these transformed variables are provided as `Xt.csv`, `Xt_subset_y.csv` and `Xt_subset_z.csv`.

6.3 DGP code

An `R` script `generate_data.R` is also provided, which implements the DGPs as described and generates the file structure describe above. At present, the code is not well formatted or documented.

NOTE: The code for generating the heteroskedastic and non-additive error data is correct, provided that the file names and flags are set by the user appropriately.

7 Evaluation: estimands and criteria

Each submitted method was evaluated according to three criteria: root mean squared estimation error on the average treatment effect *on the treated* (ATT), coverage rate of interval estimates for the ATT (for nominal 95% intervals), as well as root mean squared estimation error on the conditional average treatment effects (CATE), averaged over each observation in the sample, and average coverage of the CATEs. We also considered the length (respectively, the average length) of the reported intervals. Concretely, denote the conditional average treatment effect as

$$\tau(\mathrm{x}_i) = \mathrm{E}(Y_i^1 \mid \mathrm{x}_i) - \mathrm{E}(Y_i^0 \mid \mathrm{x}_i) = \mathrm{E}(Y_i \mid \mathrm{x}_i) - \mathrm{E}(Y_i \mid \mathrm{x}_i)$$

where the last expression follows from strong-ignorability. Denote the average treatment effect on the treated as

$$\bar{\tau}_{att} = \frac{1}{n_t} \sum_{i, Z_i = 1} \{\mathrm{E}(Y^1 \mid \mathrm{x}_i) - \mathrm{E}(Y^0 \mid \mathrm{x}_i)\} = \frac{1}{n_t} \sum_{i, Z_i = 1} \tau(\mathrm{x}_i).$$

The root mean square estimation error for the CATE is computed via

$$\mathrm{rmse}_{cate} = \sqrt{\frac{1}{n} \sum_i (\hat{\tau}(\mathrm{x}_i) - \tau(\mathrm{x}_i))^2}$$

and for the ATT it is computed as

$$\text{rmse}_{att} = \sqrt{(\hat{\tau}_{att} - \bar{\tau}_{att})^2} = |\hat{\tau}_{att} - \bar{\tau}_{att}|.$$

Note that in some publications rmse_{cate} is called the “precision in the estimated heterogeneous effects” or PEHE.

Coverage is computed as

$$\text{cover}_{att} = \frac{1}{m} \sum_j \mathbb{1}(l < \bar{\tau}_{att} < u)$$

where l and u are the lower and upper bounds of the reported interval estimate and the sum is taken over each replicate (specifically, $m = 250$ in this simulation). Likewise,

$$\text{cover}_{cate} = \frac{1}{m} \sum_j \frac{1}{n} \sum_i \mathbb{1}(l(\mathrm{x}_i)i < \tau(\mathrm{x}_i) < u(x_i)).$$

Similarly, one can consider this quantity averaged only over the treated (respectively, untreated) units:

$$\text{cover}_{catt} = \frac{1}{m} \sum_j \frac{1}{n_t} \sum_{i, Z_i = 1} \mathbb{1}(l(\mathrm{x}_i)i < \tau(\mathrm{x}_i) < u(x_i)).$$

An `R` script implementing these functions is provided as `evaluation_functions.R`.

8 Teams and partial results

Twenty-one entries were accepted and evaluated. Four of these entries were submitted by us on behalf of the ACIC conference. The first was a simple linear model for a straw man comparison. The second was a method utilizing the Bayesian additive regression tree (BART) model as implemented in the `dbarts` package; this was a high-performing entry from the 2016 version of the challenge. The latter two methods, Bayesian Causal Forests and Gradient Random Forests, were submitted by us, *ex post*, in response to results that were presented at ACIC 2017. The submission scripts for these two methods (21 and 22, below) are provided as `bcf.att.R` and `grf.att.R` respectively.

We will not describe the submitted methods in detail here; we merely provide a suggestive entry name and the names of team members.

1. Linear model
2. BART
3. Super Learner + Target Maximum Likelihood Estimation (TMLE)
4. h20 Grid
5. Propensity score regression
6. BART (multiple chains)
7. BART (symmetrized)
8. BART with propensity score
9. BART with propensity score (symmetrized)
10. BART + TMLE
11. BART + inverse probability of treatment weighting (IPTW)
12. BART + inverse probability of treatment weighting (IPTW) (symetrized)
13. Targeted Learning
14. BART with influence function
15. Super Learner
16. Sparse regression
17. X-Learner BART
18. X-Learner hRF
19. Good Cause 1
20. Good Cause 2
21. Bayesian Causal Forest[*]
22. Gradient random forest[*]

Note that method 3 does not estimate CATEs.

Because many authors submitted multiple methods, there were far fewer teams than there were methods. Using the numbering of methods from above:

- method 3: Susan Gruber and Mark van der Laan,
- method 4: Frederico Nogueira,
- method 5: Xu Shi,
- methods 6 - 12: Nicole Bohme Carnegie and Jennifer Hill,
- method 13: Chris Kennedy, Jonathan Levy, Caleb Miles, Ivana Malenica, Nima Hejazi, Andrew Kurepa Waschka and Alan Hubbard,
- method 14: Razieh Nabi Abdolyousefi, Illya Shpitser, Razieh Nabi and Alex Gain,

[*] Submitted after preliminary results were reported at ACIC 2017.

- method 15: Ryan Andrews, Illya Shpitser, Razieh Nabi and Alex Gain,
- method 16: Marc Ratkovic,
- methods 17 and 18: Soeren Kuenzel, Jasjeet Sekhon, Peter Bickel and Bin Yu,
- methods 19 and 20: Naama Parush, Chen Yanover, Yishai Shimoni and Amit Gruber.
- methods 1, 3, 21 and 22: P. Richard Hahn and Vincent Dorie

8.1 Summary plots

Summary plots are supplied as a separate PDF file named `eval_plots.R`. It contains results aggregated over the following subsets of the DGPs:

- all i.i.d. DGPs (8 total),
- all non-additive error DGPs (8 total),
- all group correlated error DPGs (8 total),
- All homoskedastic DGPs (24 total)

An `R` script generating these plots is provided as `summary_plots.R`; it pulls results data from `.Rdata` files provided in the containing folder.

8.2 Summary of findings

Because we considered a wide array of DGPs and several distinct evaluation criteria, it is difficult and also unhelpful to declare a "winner"; the word "challenge", as opposed to "contest", is specifically intended to de-emphasize the competition aspect of this exercise. Still, certain notable trends did emerge, some of which we point out here.

- 95% nominal coverage is achieved by no method. Despite the use of actual replicates, where only the noise variable changes, no method consistently covered at the desired rate.
- Methods 8, 9 and 21 — which specifically incorporate an estimate of the propensity score as covariates when estimating the response surface – did particularly well in the targeted selection regimes studied this year. While there is a long literature establishing the utility of *combining* response surface and propensity score estimates for more efficient and robust estimation of aggregate treatment effects (e.g., Bang and Robins

[2005] and Van Der Laan and Rubin [2006]), to our knowledge most methods treat each prediction problem separately. Incorporating information about the selection process into response surface estimation seems worthy of further investigation; see Hahn et al. [2017] for some initial investigations.

References

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

J. Brooks-Gunn, F. R. Liaw, and P. K. Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. 2017.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.