# Second Annual Causal Inference Data Analysis Challenge

Organized in conjunction with the 2017 Atlantic Causal Inference Conference (Conference registration is *not* required to participate.)

**Important Dates:**
May 9: Final day to submit an entry
May 23-25: Results announced at Atlantic Causal Inference Conference at UNC

**Introduction**
The causal inference challenge is an opportunity for researchers in causal inference methods to see how their preferred estimator/method stacks up against competing methods when applied to data from a range of data generating processes cooked up by a third party (us!).

The goal of the challenge is to supplement theoretical work with a wide range of simulation studies with an eye towards better understanding the operating characteristics of proposed causal inference methods in finite samples and under conditions thought to reflect common applied settings.

While many elements of the challenge are carried over from last year, new data generating processes will be used. Next year the creation of the DGP will be handed off to a new organizing committee. Over time, the objective is to compile a broad collection of data generating processes that researchers can leverage in their own methodological investigations.

**Structure of the Challenge**
Similar to last year's competition, there are 58 provided covariates (features, predictor variables), taken from a real study, which will remain fixed across all datasets. Given these covariates, the binary treatment assignment and continuous outcome will be simulated for each dataset such that ignorability (selection on observables, all confounders measured, no hidden bias…) holds. Likewise, there will be always be a sufficient common support for the treatment group to estimate the treatment effect on the treated.

The simulated treatment and response pairs will be simulated from 32 distinct data generating processes (DGP), which differ from one another in various respects (we're not telling which, that would ruin the fun). For each DGP, 250 replicate datasets will be produced, to aid in studying coverage (for a total of 8000 total datasets).

**How to participate:**

1. Create an executable or script. We support the following languages: R, Stata, Matlab, Python, C++.
2. The data file will be in comma-separated value (csv) format and match the following specification:
    - Column 1 is a binary treatment variable
    - Column 2 is a continuous response variable
    - Columns 3 and above are covariates; categorical variables/factors are coded with letters A/B/C/…, binary variables are 0/1, and other columns are real numbers
3. Your executable should take three inputs: the name of an input data file and the names of two output files.
4. Your executable should create two outputs:
    - The first output should consist of a 3-column csv file containing the estimate of the sample average treatment effect on the treated, and lower and upper bounds for a 95% confidence interval.
    - The second output should be a csv file with individual causal estimates, one per row in the same format as above (three columns, point estimate, lower bound, upper bound).
5. An example in R including test data and output is available here.
6. Submit your script by email to vdorie@gmail.com with subject line “causal inference challenge 2017”.

**Evaluation of Method Performance**

The estimand of interest will be the Sample Average effect of the Treatment on the Treated (SATT). That is, if we let Z denote binary treatment assignment and Y(0), Y(1) denote the continuous potential outcomes with respect to that treatment, the estimand of interest is E[Y(1)-Y(0) | Z=1] where the expectation is taken over the sample. We will additionally consider individual treatment effects for individuals in the sample.

We will evaluate estimators based on several criteria: root mean squared estimation error of the sample average treatment effect on the treated, the corresponding bias, coverage, confidence interval length, computational time, and average root mean squared estimation error of the individual treatment effects.

**Dissemination of Results**

Challenge results will be revealed at the conference. In addition, a manuscript will be prepared describing the details of the data generating processes and the evaluation of the results. After

that manuscript has been accepted for publication an R package will be released that will allow researchers to replicate the simulated data.

**Last Year's Results**
Finally, for interested parties we have posted solutions to the 20 Do-It-Yourself simulations from last year. Note that the data-generating-processes is completely different this year so that while the data can provide a way to calibrate a method for these covariates, they may also lead to overfitting.

**Challenge Organizers:**
Richard Hahn, Vincent Dorie, and Jared Murray. (Note: challenge organizers are not permitted to submit entries.)

**Questions:**
If there are aspects of the competition that are as yet still unclear please feel free to contact us at richard.hahn@chicagobooth.edu.