2017 Statistics Methods Forum Data Challenge

Eric Polley

Nov. 15th, 2017

Introduction

The focus this year is the estimation of a causal treatment effect from a retrospective study[1]

A dataset with 400 patients will be provided with a binary treatment, a continuous outcome of interest, and a set of potential confounders or covariates. The primary goal is to estimate the average treatment effect and provide a 95% confidence interval for the estimate.

Details for the Challenge available on Github:
https://github.com/ecpolley/Data_Challenge_2017

[1] Partially motivated by the Atlantic Causal Inference Data Challenge http://causal.unc.edu/acic2017/

Outline

Will continue for the next two regular Statistical Methods Forums (2-3pm central)

- Oct. 18th: Introduction to the data challenge and dataset
- Nov. 15th: Group discussion and Q&A session
- Dec. 18th, 5:00pm local: Team submissions deadline (If team is across sites, depends who sends the results)
- Dec. 20th: Final results and team scores, and discussion of methods used

Team Science

- Only 1 team sent a name (The Significant Six)
- Please let me know if you have a team working on the project (email: Polley.Eric@Mayo.edu), we won't hold you to submit results, but helpful for us to know who is working on the challenge.

Primary Objective

The primary goal is to estimate the average treatment effect (ATE). We can define the values $Y(0)$ and $Y(1)$ to be the possibly counterfactual outcome values had the patient been given treatment 0 and treatment 1, respectively. In the dataset, the observed value Y is:

$$Y_i = (1 - A_i)\,Y_i(0) + A_i\,Y_i(1)$$

The parameter of interest is the ATE:

$$\psi = E(Y(1) - Y(0))$$

and provide a 95% confidence interval for the estimate.

Teams scores based on distance between estimate and true value, and the width of teh confidence interval. A penalty will be added if the true value is outside the interval.

Primary Objective

Team results can be emailed to Eric (Polley.Eric@Mayo.edu), with the following:

1. Team members
2. Team name
3. ATE estimate
4. Lower and Upper confidence limits

Secondary Objective

The secondary goal is to estimate the individual treatment effect for all 400 samples:

$$\psi_i = Y_i(1) - Y_i(0), \ i \in 1, \ldots, N$$

The mean squared error with the true individual treatment effect will be computed (*i.e.* precision in estimation of heterogeneous effects), along with the concordance of the sign $(+/-)$ of the effect.

Secondary Objective

Team results for the optional secondary objective can be emailed to Eric (Polley.Eric@Mayo.edu) with the following:

1. Team members
2. Team name
3. Text file with 2 columns: ID variable and predicted individual treatment effect

Questions?