Introduction to 2017 Statistics Methods Forum Data Challenge

Eric Polley

Oct. 18th, 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

- Participants are encouraged to work in teams ($N \in (1, 2, \ldots, 10)$)
- Opportunity to learn from each other and work with people outside usual team
- Data is publicly available, so is available outside Mayo
- If you would like help forming a team, email Eric Polley, Kristin Mara, or Sara Fett
- Teams are responsible for creating a team name, and may submit up to 3 estimates, with the last submission being the official one
- If you are participating, please let us know in case we have any Data Challenge announcements

Overview of Dataset

- CSV file available on Github, `https://github.com/ecpolley/Data_Challenge_2017`
- 400 independent observations (rows)
- Y: Continuous outcome of interest
- A: Binary treatment indicator
- W1-W25: Baseline measured variables, includes all confounders
- ID: Individual id number
- No missing data, No hidden messages in 10-dimension space[2]

[2] See `https://github.com/ecpolley/CSMF_Data_Challenge/blob/master/Biomarkers.R`

Overview of Dataset

```r
# link to data on GitHub page if not available
if(file.exists("Data.csv")) {
  Dat <- read.csv("Data.csv")
} else {
  urlfile <- "https://raw.githubusercontent.com/ecpolley/
    Data_Challenge_2017/master/Data.csv"
  download.file(urlfile, destfile = "Data.csv")
  Dat <- read.csv("Data.csv")
}
dim(Dat)
```

```
## [1] 400  29
```

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:

$$Y_i = (1 - A_i)Y_i(0) + A_iY_i(1)$$

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

Example (Ignore Confounding)

```
t.test(Y~A, data = Dat)
```

```
##
##  Welch Two Sample t-test
##
## data:  Y by A
## t = -11.134, df = 342.67, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -4.040489 -2.827246
## sample estimates:
## mean in group 0 mean in group 1
##        4.667309        8.101176
```

```
with(Dat, mean(Y[A == 1]) - mean(Y[A == 0]))
```

```
## [1] 3.433867
```

Extensions

- As a team should discuss different methods to adjust for confounding (*e.g.* Regression model, Machine Learning, IPTW, propensity score matching, Targeted MLE, etc.)
- Depending on selection of method for estimation, need to estimate confidence interval (*e.g.* closed form approximation, resampling, etc.)

Questions?