Data Science for Causal Inference Lab

Ryan T. Moore[*]

2024-07-15

1 Causal Forests

Consider the social pressure experiment data from Gerber, Green, and Larimer (2008). Read in the data, prepare it for processing with `glmnet` and `grf`, and take a sample of 50,000 of the registrants (some code provided below to help). We are interested in treatment effect heterogeneity surrounding the causal effect of the "neighbors" message on turnout in the 2006 presidential primary.

Build an honest causal forest using the predictors `age`, `hhsize`, `isFemale`, and `primary2004`. Examine the treatment effect heterogeneity along the dimensions of the predictors.

```
set.seed(233559574)
social <- read_csv("http://j.mp/2Et71U0")

social <- social |>
  mutate(age = 2006 - yearofbirth,
         isFemale = (sex == "female"),
         isFemale = as.numeric(isFemale),
         sentNeighbors = (messages == "Neighbors"),
         sentNeighbors = as.numeric(sentNeighbors))

social <- social |> sample_n(50000)
```

2 Variable Selection

Starting with the same data, do some naïve "feature engineering" by adding age^2, age^3, age^4, and age^5 to the matrix `X`. Use 10-fold cross-validation on the LASSO, and report the coefficients associated with (a) the λ that minimises the mean cross-validated MSE, and (b) the λ that gets within 1 SE of the mean cross-validated MSE. Describe any differences in which coefficients are retained, and their magnitudes. Finally, what experimental treatment effects do you estimate using these two different λ values?

[*] Department of Government, American University, Kerwin Hall 228, 4400 Massachusetts Avenue NW, Washington DC 20016-8130. +1 202.885.6470 (tel); `rtm` (at) `american` (dot) `edu`; http://ryantmoore.org.

3 Variable Selection

Consider a higher-dimensional observational dataset from your research. Use a double LASSO to select variables and estimate an ATE of interest.

References

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.