Statistical Learning in Mediation Analysis

Lab for Chapter 6: Estimating direct and indirect effects for stochastic interventions in R

David Benkeser
Emory University

Iván Díaz
New York University

Marco Carone
University of Washington

MODULE 13

Summer Institute in Statistics for Clinical and Epidemiological Research

July 2023

Contents of this lab

1. Illustration of estimation of the direct and indirect effects using the `medshift` R package

Illustrative dataset

- Recall the dataset `weight_behavior` from the mma R package used in Lab 3.
- We set up the dataset the same way as for Lab 3, removing missing data:

```
library(mma)
library(tidyverse)
# load and examine data
data(weight_behavior)
dim(weight_behavior)
```

```
## [1] 691  15
```

```
# drop missing values
weight_behavior <- weight_behavior %>%
  drop_na() %>%
  as_tibble()
```

Setting up the problem

As in Lab 3, we focus on the causal effects of participating in a sports team (`sports`) on the BMI of children (`bmi`), taking into consideration mediators given by (`snack`, `exercises`, `overweigh`). All other measured covariates are taken to be potential baseline confounders.

Instead of measuring the effect of a binary intervention intervening on participation in a sports team, we conceptualize the question in terms of increasing the likelihood of participation.

For this, we use an incremental propensity score intervention where we wonder what would have happened if the odds of participating would have been 2 times higher than they actually were.

The function `medshift()`

The package may be installed running

```
library(devtools)
install_github('nhejazi/medshift')
```

The main function of the package is `medshift()`. The main arguments are as follows:

- `W`: a data frame with baseline confounders
- `A`: a binary (zero or one) treatment variable
- `Z`: a mediator of interest
- `delta`: the incremental odds ratio
- `Y`: binary or continuous outcome vector
- `g_learners`: an sl3 learner stack for $P(A = a \mid W = w)$
- `e_learners`: an sl3 learner stack for $P(A = a \mid Z = z, W = w)$
- `m_learners`: an sl3 learner stack for $E(Y \mid A = a, Z = z, W = w)$
- `estimator`: which estimator is to be used "onestep" or "tmle"
- `estimator_args`: other estimation parameters such as the number of cross-fitting folds

The function `medshift()`

Note two quirks of the function `medshift()`:

- The mediator of interest is denoted Z, not M (this is due to notational differences in the original research articles where these methods were proposed)
- The function `medshift()` does not directly estimate direct or indirect effects. Instead, it estimates the parameter $E[Y(A_\delta, M)]$ which constitutes the building block for mediation (see main chapter slides)
- The other parameters for mediation, namely $E[Y(A_\delta)]$ and $E[Y]$ may be estimated using the `ipsi` function (as illustrated in the main chapter) and the empirical mean, respectively.

Estimating the IPSI direct effect

First, we set up the super learner libraries as in Lab 3:

```
library(sl3)
# instantiate learners
fglm_lrnr <- Lrnr_glm_fast$new()
lasso_lrnr <- Lrnr_glmnet$new(alpha = 1, nfolds = 3)
rf_lrnr <- Lrnr_ranger$new(num.trees = 200)
# create learner library and instantiate super learner ensemble
lrnr_lib <- Stack$new(fglm_lrnr, lasso_lrnr, rf_lrnr)
sl_lrnr <- Lrnr_sl$new(learners = lrnr_lib, metalearner = Lrnr_nnls$new())
```

Estimating the IPSI direct effect

```r
stoch_decomp_onestep <- medshift(
  W = weight_behavior[, c("age", "sex", "race", "tvhours")],
  A = (as.numeric(weight_behavior$sports) - 1),
  Z = weight_behavior[, c("snack", "exercises", "overweigh")],
  Y = weight_behavior$bmi,
  delta = 2,
  g_learners = lasso_lrnr,
  e_learners = lasso_lrnr,
  m_learners = lasso_lrnr,
  estimator = "onestep",
  estimator_args = list(cv_folds = 5)
)
summary(stoch_decomp_onestep)
```

```
##       lwr_ci    param_est      upr_ci    param_var
##     18.74992    19.078205    19.40649     0.028055
##     eif_mean    estimator
## 7.236053e-16      onestep
```

This gives us $E[Y(A_\delta, M)]$. We will now contrast it with $E(Y)$.

Estimating the IPSI direct effect

First, we create a convenience function

```
linear_contrast <- function(params, eifs, ci_level = 0.95) {
  # bounds for confidence interval
  ci_norm_bounds <- c(-1, 1) * abs(stats::qnorm(p = (1 - ci_level) / 2))
  param_est <- params[[1]] - params[[2]]
  eif <- eifs[[1]] - eifs[[2]]
  se_eif <- sqrt(var(eif) / length(eif))
  param_ci <- param_est + ci_norm_bounds * se_eif
  # parameter and inference
  out <- c(param_ci[1], param_est, param_ci[2])
  names(out) <- c("lwr_ci", "param_est", "upr_ci")
  return(out)
}
```

Results

```r
EY <- mean(weight_behavior$bmi)
eif_EY <- weight_behavior$bmi - EY
params_de <- list(stoch_decomp_onestep$theta, EY)
eifs_de <- list(stoch_decomp_onestep$eif, eif_EY)

# direct effect = EY - estimated quantity
de_est <- linear_contrast(params_de, eifs_de)
de_est
```

```
##       lwr_ci   param_est      upr_ci
## -0.50912890 -0.04890266  0.41132358
```

From this we can conclude that increasing the odds of participation in a sports by 2 leads to a relatively small direct effect on BMI. (More complete conclusions would require estimating also the total effect.)