

Statistical Learning in Mediation Analysis

Lab for Chapter 4: Estimating controlled direct effects in R

David Benkeser
Emory University

Iván Díaz
New York University

Marco Carone
University of Washington

MODULE 14

**Summer Institute in Statistics for
Clinical and Epidemiological Research**
July 2024

Contents of this lab

- 1 Illustration of estimation of the controlled direct effect using the [lmtpl R package](#)
- 2 [lmtpl](#) integrates Super Learning via the implementation in the [sl3 R package](#) as well as the [SuperLearner R package](#)
- 3 To learn more about [lmtpl](#), check out the package's vignette
- 4 The same effect could also have been estimated using the [ltmle R package](#), with the only difference that [ltmle](#) does not implement cross-fitting.

Illustrative dataset

- We re-analyze the data from a recent study examining gender differences in wage expectations among students at two Swiss institutions of higher education (Fernandes et al., 2021).
- We study the causal relation between gender (A) and wage expectations three years after graduation (Y) study program (Z) and whether a student plans to continue obtaining further education or work full time after graduation (M) as mediators.
- The dataset is freely available in the [causalweight](#) R package

Setting up the dataset

```
data(wexpect)
W <- wexpect %>%
  dplyr::select(age, swiss, hassiblings, motherhighedu, fatherhighedu,
               motherworkedfull, motherworkedpart,
               matwellbeing, homeowner, treatmentinformation)
W <- data.frame(model.matrix( ~ 0 + ., W))
colnames(W) <- paste0('W', 1:ncol(W))
A <- pull(wexpect, male)
M <- pull(wexpect, plansfull)
Z <- wexpect %>% select(business, econ, communi, businform)
colnames(Z) <- paste0('Z', 1:ncol(Z))
Y <- pull(wexpect, wexpect2)
data <- data.frame(W, A, Z, M, Y)
```

Setting up the learners

The CRAN version of `lmtp` uses the `SuperLearner` package for estimation of nuisance parameters. Therefore, specification of the `SuperLearner` libraries follows the same syntax:

```
sl_lib <- c('SL.glm', 'SL.ranger', 'SL.earth')
```

For this illustrative example we have included only the above three libraries, but in applications you should include more!

The main functions in `lmtp`

- The package implements estimators of very general effects for longitudinal data, including modified treatment policies, dynamic treatment regimes, and static regimes
- As seen in this workshop, the controlled direct effect under intermediate confounding is a particular case of a static regime with two time points
- The `lmtp` package has two main functions: `lmtp_sdr()` and `lmtp_tmle()`
- The SDR estimator is more robust to model misspecification compared to the TMLE at the expense of not being a substitution estimator

Using `lmtpr_sdr()` and `lmtpr_tmle()` for the CDE

These are the relevant function arguments for mediation analysis:

- `data`: a data frame with all the relevant variables
- `trt`: a vector containing the names of the intervention nodes (for mediation, this is the treatment and the mediator)
- `outcome`: the column name for the outcome variable
- `baseline`: a vector of names of baseline variables
- `time_vary`: a list of names of time-varying variables (for mediation, this is the intermediate confounders Z)
- `shift`: a function specifying the intervention of interest
- `intervention_type`: 'static' for mediation analysis
- `outcome_type`: 'continuous' or 'binomial'
- `learners_outcome`: learners for the outcome model
- `learners_trt`: learners for the treatment/mediator model

Estimating the controlled direct effect

First, we create two shift functions, one for estimating $E[Y(1,0)]$, and another for $E[Y(0,0)]$.

```
shift10 <- function(data, trt) {  
  if(trt == 'A') return(rep(1, length(data[[trt]])))  
  if(trt == 'M') return(rep(0, length(data[[trt]])))  
}  
shift00 <- function(data, trt) {  
  if(trt == 'A') return(rep(0, length(data[[trt]])))  
  if(trt == 'M') return(rep(0, length(data[[trt]])))  
}
```

Estimating the controlled direct effect

Now we call the estimating function with the arguments `shift10` and `shift00`

```
EY10 <- lmtp_sdr(data, trt = c('A', 'M'), outcome = 'Y', baseline = names(W),  
                 time_vary = list(NULL, names(Z)), shift = shift10,  
                 learners_outcome = sl_lib, learners_trt = sl_lib,  
                 intervention_type = 'static', outcome_type = 'continuous')  
EY00 <- lmtp_sdr(data, trt = c('A', 'M'), outcome = 'Y', baseline = names(W),  
                 time_vary = list(NULL, names(Z)), shift = shift00,  
                 learners_outcome = sl_lib, learners_trt = sl_lib,  
                 intervention_type = 'static', outcome_type = 'continuous')
```

Estimating the controlled direct effect

The function `lmtptest` provides a convenient way to construct a confidence interval for the CDE:

```
contrast <- lmtptest(EY10, ref = EY00)
contrast

##
##
##   theta shift  ref std.error conf.low conf.high p.value
## 1   1.35  10.1  8.78    0.0389     1.28     1.43 <0.001
```

Here we are setting the mediator to $M = 0$, which means that we are estimating the effect of gender in a hypothetical world where all students plan to continue obtaining education after graduation.

We conclude that the effect of gender on wage expectations in a hypothetical world where everyone continues education after graduation is of approximately $1.35 \times 500 = 675.92$ CHF gross per month (see documentation of the data for the coding of the outcome that allows this interpretation).

Do-it-yourself analysis of a real dataset

- 1 For this example we revisit the dataset `framing` available in the `mediation` R package considered in Lab 3.

- 2 Recall the question answered in Lab 3:

To what extent is the causal effect of the tone of the story on negative attitude towards immigration mediated by anxiety?

- 3 Consider the potential intermediate confounder `p_harm`, which denotes a subjects' perceived harm caused by increased immigration.
- 4 Dichotomize the anxiety variable by considering two categories: 'not anxious at all' vs the rest
- 5 Estimate the controlled direct effect $E[Y(1, 0) - Y(0, 0)]$ comparing hypothetical worlds where the tone of the story was positive vs negative, fixing anxiety to 'not anxious at all', adjusting for the intermediate confounder `p_harm`
- 6 Contrast this with the results of the previous Lab 3