Day 3, Practical 1

Helene Charlotte Wiese Rytgaard

September 29, 2021

This practical consists of understanding the study and the results of Kreif et al. (2017). You should read the items listed below in Section 1 and write up your responses.

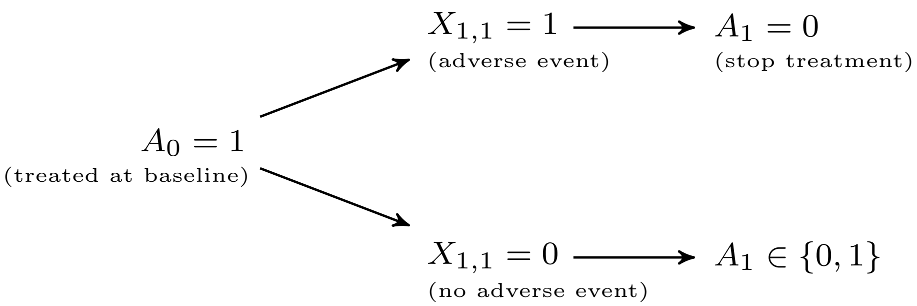
If you have more time, there is an **optional** second part of this practical (see Section 2) in which you will reproduce the simulation study of the lecture slides.

1 Questions for Kreif et al. (2017)

1. What is the overall research question considered in the paper?
2. What are the observed variables? How many longitudinal time-points are there and how are they defined?
   - e.g., what do A_t, Z_t and Y_t measure?
3. Which variables affect each other? Are there any time-dependent confounders?
4. What are the hypothetical regimes that are considered? Static regimes? Dynamic? What are the corresponding causal parameters?
5. Go over the stated causal assumptions required to identify the causal parameters. Is positivity equally likely to hold for all regimes/strategies considered?
6. Results: The authors consider a 'naive' approach, an IPW estimator, a g-formula estimator and a TMLE estimator. What results do they find with each method (see Figure 1 of the paper)? How do they differ in terms of bias and variance?

2 Optional: Reproducing the simulation study of the lecture slides

We consider data simulated as follows: $X_{0,1}, X_{0,2}, X_{0,3}$ are baseline covariates, $A_0 \in \{0,1\}$ is a randomized treatment indicator, $X_{1,1}, X_{1,2}$ are follow-up covariates, $A_1 \in \{0,1\}$ is a follow-up treatment decision, and $Y \in \{0,1\}$ is the final outcome. We can think of the variable $X_{1,1}$ as an indicator of an adverse event from the baseline treatment, an adverse event that causes treated subjects to switch from 'treatment' ($A_0 = 1$) to 'no treatment' ($A_1 = 0$) — see the figure below. We can further think of the variable $X_{1,2}$ as a marker of being likely to forget to take the medicin (or thinking it is too bothersome) which increases the probability of switching treatment as well.



Particularly, the data are simulated (with sample size `n`) as follows:

```r
# baseline covariates
X0.1 <- runif(n, -2, 2)
X0.2 <- rnorm(n)
X0.3 <- rbinom(n, 1, 0.2)

# baseline treatment (randomized)
A0 <- rbinom(n, 1, 0.5)

# follow-up covariates
X1.1 <- rbinom(n, 1, plogis(-0.7 + 0.3*X0.3 + 0.8*A0))
X1.2 <- rbinom(n, 1, plogis(0.25 - 0.55*X0.3))

# follow-up treatment
A1 <- rbinom(n, 1, prob=plogis(0.9 - 5*(1-A0) - 4.7*X1.1 - 4.8*X1.2))

# outcome
Y <-  rbinom(n, 1, prob=plogis(-0.9 - 0.2*A0 + 1.2*X1.1 - 0.1*A1 - 0.8*A1*(X1.1==0)))
```

Task 1: Write a function that takes n as input and returns the observed data in a data frame.

We are interested in the effects of different types of interventions:

1. The intention-to-treat (ITT) effect which only intervenes on treatment at baseline and contrasts the two scenarios of being treated at baseline ($A_0 = 1$) and not being treated at baseline ($A_0 = 0$).
2. The (static) effect of being 'always treated' ($A_0 = A_1 = 1$) contrasted to 'never treated' ($A_0 = A_1 = 0$).

3. A dynamic effect of being treated at baseline ($A_0 = 1$) and only treated at follow-up if the adverse event has not happened, i.e., $X_{1,1} = 0$ — contrasted to being 'never treated' ($A_0 = A_1 = 0$).

Task 2: Update your simulation function from **Task 1** so that it takes as argument the choice of one of the effects 1.–3. above and returns (an approximation to) the true value of that effects (the function should still give as output the observed data when no intervention is specified). Use the function to compute the true values of each parameter.

Task 3: In this task we focus on the effect of the static interventions (2. above), taking a naive approach to estimation. First, simulate a dataset with sample size $n = 1000$ using the function from **Tasks 1/2**. Then, specify a multivariate logistic regression of the outcome regressed on all treatment variables and all covariates. Get means of the predictions under $A_0 = A_1 = 1$ and contrast it to the mean of the predictions under $A_0 = A_1 = 0$. Does this correctly estimate the static effect? Next, specify a multivariate logistic regression of the outcome regressed on both treatment variables and baseline covariates (leaving out follow-up covariates!). Get means of the predictions under $A_0 = A_1 = 1$ and contrast it to the mean of the predictions under $A_0 = A_1 = 0$. Does this correctly estimate the static effect?

Task 4: In this task, you should make a simulation study out of **Task 3**. In each repetition, draw a new dataset from your simulation function, and then construct the two naive estimators as in **Task 3**. Repeat the simulations 500 times and save the estimates. Plot the histogram and mark the true value by a vertical line. What do you see?

References

Kreif, N., L. Tran, R. Grieve, B. De Stavola, R. C. Tasker, and M. Petersen (2017). Estimating the comparative effectiveness of feeding interventions in the pediatric intensive care unit: a demonstration of longitudinal targeted maximum likelihood estimation. *American journal of epidemiology 186*(12), 1370–1379.