Day 3, Practical 2

Helene Charlotte Wiese Rytgaard

June 8, 2023

In this practical we consider various versions of longitudinal settings, to define dynamic interventions and illustrate issues arising in presence of death. The goal is to get acquainted with the longitudinal data structure, as well as the benefits and possibilities in using a causal framework to define target parameters. Specifically, we will use simulation studies to:

1. Demonstrate the distinctions between various static and dynamic treatment interventions, and the immediate challenges for estimation posed by time-dependent confounding.
2. Demonstrate interpretational challenges associated with targeting hazard ratios in survival analysis, and the benefits of targeting parameters defined under dynamic interventions.

The following provides an overview of the different parts of the practical:

1. In Section 1 we consider a simulation setting with a longitudinal treatment and different (simple) versions of static/dynamic interventions. This simulation setting will be used again tomorrow, where we introduce the TMLE algorithm for longitudinal data structures.
2. In Section 2 we consider a simple (discrete) survival setting with just two time-points. Here we consider effects on the hazard versus the absolute risk scale, with focus on three different variations of the setting:
   - baseline treatment (and unobserved heterogeneity).
   - (longitudinal) treatment switching.
   - baseline treatment and right-censoring.

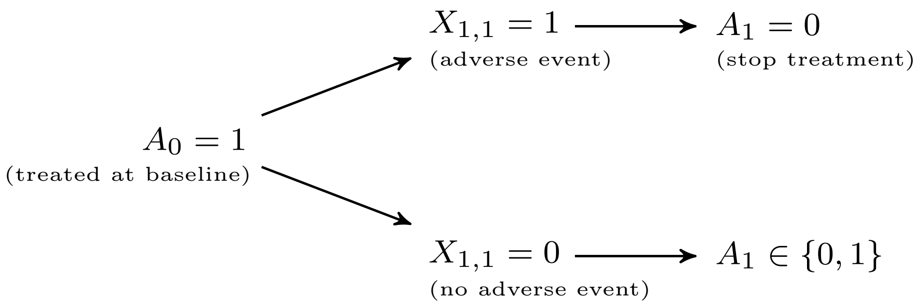
Note that you probably do not have time to finish all parts.

1 Simulation setting 1

In the first part of this practical, we consider longitudinal 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;
- $Y \in \{0, 1\}$ 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.



Task 1: Write a function with argument `n` so that you can simulate observed data with a given sample size (`n`) such that:

1. $X_{0,1}$ is uniform on $[-2, 2]$.
2. $X_{0,2}$ follows a normal distribution with mean 0 and variance 1.
3. $X_{0,3}$ is a binomial variable with $P(X_3 = 1) = 0.2$.
4. A_0 is randomized with $P(A_0 = 1) = 0.5$.
5. The distribution of $X_{1,1}$ is given by the following logistic regression model:
$$\mathbb{E}[X_{1,1} \mid X_{0,1}, X_{0,2}, X_{0,3}, A_0] = \text{expit}(-0.7 + 0.3X_{0,3} + 0.8A_0).$$
6. The distribution of $X_{1,2}$ is given by the following logistic regression model:
$$\mathbb{E}[X_{1,2} \mid X_{0,1}, X_{0,2}, X_{0,3}, A_0] = \text{expit}(0.25 - 0.55X_{0,3}).$$
7. The distribution of A_1 is given by the following logistic regression model:
$$\mathbb{E}[A_1 \mid X_{0,1}, X_{0,2}, X_{0,3}, A_0, X_{1,1}, X_{1,2}] = \text{expit}(0.9 - 5(1 - A_0) - 4.7X_{1,1} - 4.8X_{1,2}).$$
8. The distribution of Y is given by the following logistic regression model:
$$\mathbb{E}[Y \mid X_{0,1}, X_{0,2}, X_{0,3}, A_0, X_{1,1}, X_{1,2}, A_1] = \text{expit}(-0.9 - 0.2A_0 + 1.2X_{1,1} - 0.1A_1 - 0.8A_1(1 - X_{1,1})).$$

1.1 Effects of interventions

We are interested in the causal risk difference constrasting different types of interventions:

1. The intention-to-treat (ITT) effect which is defined under interventions only 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.

1.2 "Naive" estimation

Task 3: In this task we focus on the effect of the static interventions from Section 1.1, taking a naive approach to estimation. Follow the steps below:

1. First, simulate a dataset with sample size $n = 1000$ using the function from **Task 1** and **Task 2**.

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?

3. 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?

2 Simulation setting 2

In this simulation setting we consider data as illustrated in Figure 1. Note that this is a "discrete-time" survival setting, where hazards are defined as

$$\lambda(t \mid A) = P(T = t \mid T \geq t, A).$$

If you are familiar with survival analysis, note that we consider here effects from logistic regression models rather than Cox proportional hazards models (only to make this practical simpler).

Task 5: Write a function with argument `n` so that you can simulate observed data with a given sample size (`n`) such that:

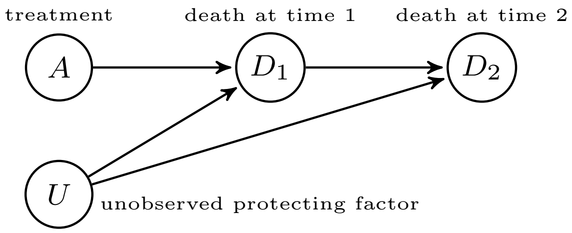


Figure 1

1. U is a binary variable with $P(U = 1) = 0.5$.
2. A is a binary variable with $P(A = 1) = 0.5$.
3. D_1 is binary survival status with distribution given by the following logistic regression model:
$$\mathbb{E}[D_1 \mid A, U] = \text{expit}(1.3 - 1.8U - 1.1A).$$
4. $D_2 = 1\{D_1 = 0\}\tilde{D}_2$, where the distribution of $\tilde{D}_2$ is given by the following logistic regression model:
$$\mathbb{E}[\tilde{D}_2 \mid A, U] = \text{expit}(2.1 - 3.9U).$$

2.1 Interpretation of hazard (odds) ratios

Comment. The point of the following task is to illustrate that the hazard (or here, odds) cannot be used to say anything about the direct effect of exposure on mortality at the second time-point. The problem is that it is defined conditional on being alive at the first time-point; subjects alive at the second time-point are different due to the unobserved protecting factor.

Task 6: Draw a single dataset of sample size $n = 1e6$ from your function from **Task 5**. For this data, fit a model for the period-1 specific hazard and the period-2 specific hazards as follows:

1. Fit a logistic regression with D_1 as outcome and treatment A as the only predictor. This yields a model for $\lambda(t_1 \mid A)$. Comment on the estimated coefficient for the effect of A.
2. Fit a logistic regression with D_2 as outcome *among those alive at time 1* (i.e., $D_1 = 0$) and treatment A as the only predictor. This yields a model for $\lambda(t_2 \mid A)$. Comment on the estimated coefficient for the effect of A.

3. Make a long version of the dataset where a single binary variable D keeps track of survival status for each individual, and where each individual contributes with a row as low as they are alive. Further add a factor variable `t` $\in \{1, 2\}$ that indicates the time period. This should look about as follows:

```
   id A D t
1:  1 0 0 1
2:  1 0 0 2
3:  2 0 1 1
4:  3 1 0 1
5:  3 1 0 2
6:  4 1 1 1
```

4. Run the two logistic regressions below on the stacked data from 3, and comment. Note that the first one yields a model for $\lambda(t \mid A)$, assuming a constant effect of A. How does the output from the second model the estimated coefficients in 1. and 2.?

```
glm(D~A+t, data=dat.stacked, family=binomial)
glm(D~A*t, data=dat.stacked, family=binomial)
```

5. Simulate the counterfactual outcome variables $D_1^{a_0}$, $D_2^{a_0}$ for $a_0 = 0, 1$ (corresponding to the counterfactual scenarios where $A = 0, 1$) and then the counterfactual risk difference $\mathbb{E}[D_2^1 - D_2^0]$.

6. Repeat 1.–6. with a different random seed (using `set.seed()`) to see that the estimated coefficients do not change much.

2.2 Treatment switching

Comment. The point of the following tasks is to illustrate the importance of taking death into account when defining treatment regimes over time. The "mistake" causing problems in these tasks may seem so immediately wrong that it would never be done in practice — but there are examples of publications making exactly this type of mistakes.

Task 7: To your simulation function from **Task 5**, add a variable A_1 representing follow-up treatment, and a variable $A_{\text{switch}} = 1\{A_1 \neq A\}$ representing treatment switching (if $A_1 \neq A$ for a given subject, that subject has switched treatment). Simulate $A_1 = 1\{D_1 = 0\}\tilde{A}_1 + 1\{D_1 = 1\}A$, where the distribution of $\tilde{A}_1$ is given by the following logistic regression model:

$$\mathbb{E}[\tilde{A}_1 \mid A, U] = \text{expit}(1.2 + 0.5A).$$

Task 8: Draw a single dataset of sample size $n = 1e6$ from your function from **Task 7**. Fit a logistic regression on the subset of the dataset corresponding to those with treated at baseline ($A = 1$), with D_2 as outcome and the variable A_{switch} as predictor. Comment on the estimated coefficient for the effect of A_{switch}.

Task 9: Simulate the counterfactual outcome variables $D_1^{a_0}$, $D_2^{a_0,a_1}$ for $a_0 = 1$ and $a_1 = 0, 1$, corresponding to the interventions $A = a_0$ and $A_1 = 1\{D_1^{a_0} = 0\}a_1 + 1\{D_1^{a_0} = 1\}a_0$. Compute the counterfactual risk difference $\mathbb{E}[D_2^{1,1} - D_2^{1,0}]$.

2.3 Right-censoring and non-proportionality

Comment. The point of the following tasks is to illustrate an additional problem with targeting the (associational) hazard ratio (or, here odds ratio). Indeed, introducing an independent censoring variable in the simulation setting from **Task 5** may change the "true average effect" of A when the effect of A is not the same across the time-points.

Task 10: We now extend the simulation function from **Task 5** by adding a censoring variable C_1. Particularly, you should make a new simulation function where:

1. U is a binary variable with $P(U = 1) = 0.5$.
2. A is a binary variable with $P(A = 1) = 0.5$.
3. D_1 is binary survival status with distribution given by the following logistic regression model:
$$\mathbb{E}[D_1 \mid A, U] = \text{expit}(1.3 - 1.8U - 1.1A).$$
4. $C_1 = 1\{D_1 = 0\}\tilde{C}_1$ is a binary censoring indicator variable, with distribution of $\tilde{C}_1$ given by the following logistic regression model:
$$\mathbb{E}[\tilde{C}_1 \mid A, U] = \text{expit}(1.3).$$
5. $D_2 = 1\{D_1 = 0, C_1 = 0\}\tilde{D}_2$, where the distribution of $\tilde{D}_2$ is given by the following logistic regression model:
$$\mathbb{E}[\tilde{D}_2 \mid A, U] = \text{expit}(2.1 - 3.9U).$$

Task 11: Repeat **Task 6** with the new simulation function (NB: for step 5., the counterfactuals are defined further under "no censoring"). Compare to the results you got in **Task 6**.