Outline

Writing an introduction

Parameter versus estimator versus estimate

What is the point with the simulation study?

Final data analysis

(YSD event – advertisement!)

Writing an introduction

Writing an introduction

- What is the purpose/goal of the report?

  [Both from an applied and a methodological perspective]

1. Explore estimation methods based on the efficient influence function for causal effect estimation in a particular real life problem
2. Assess if there is a causal effect of a planned cesarian section on the risk of postpartum haemorrhage

Writing an introduction

- What is the methodological approach taken?
  - We are investigating targeted learning methods for causal effect estimation in a setting where the target parameter (specifically the ATE) is low-dimensional but estimation has to deal with potentially high-dimensional nuisance parameters
  - The ATE is identified in terms of the g-formula and can be estimated with a simple two-step procedure
  - We consider different variants of constructing an estimator for the ATE and compare their performance

Writing an introduction

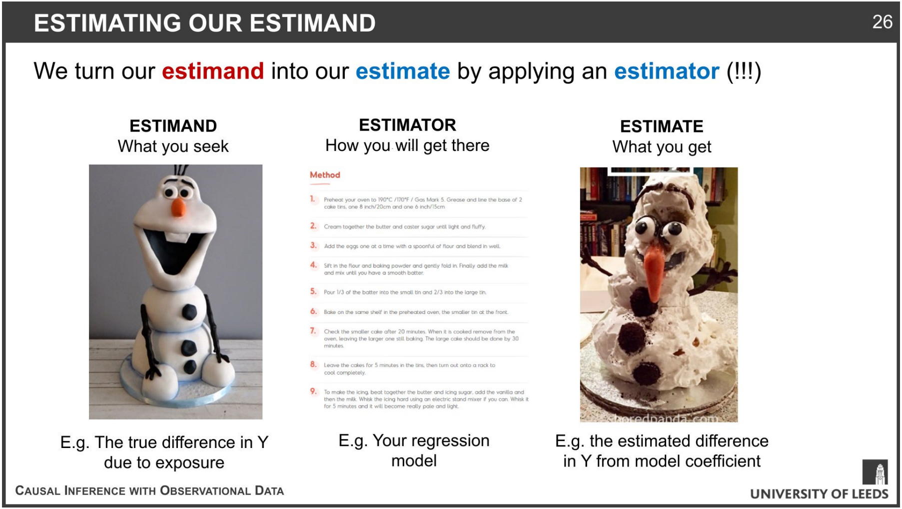
- Why are we interested in these methods and how does the report fit into the bigger picture?
  - Parametric models are quite restrictive and they require correct specification for valid statistical inference
  - We explore efficient influence function based estimation which can be used to either
    1. improve robustness of estimation based on parametric models (by using information from *both* the treatment mechanism and the outcome regression), or
    2. provide a basis for combining machine learning techniques (much more flexible than parametric models) with valid statistical inference

Writing an introduction

- What is done in the report?
  - The causal problem is clarified and translated to a statistical estimation problem
  - Tools from semiparametric efficiency theory allows us to analyze the statistical estimation problem and characterize an optimal estimator in terms of the efficient influence function
  - ...
  - Does it matter if we use efficient influence function based estimation?
  - Should we just focus on picking a good nuisance parameter estimator?
  - ...
  - To explore and assess different estimators, the full data is divided into two parts:
    1. One part is used for exploration/simulation studies.
    2. The other part is used for the final analysis

Parameter versus estimator versus estimate

Parameter versus estimator versus estimate



Parameter versus estimator versus estimate

- The parameter represents what we want to estimate
  - for our purposes, there is only one parameter (the ATE)
- The estimator is an algorithm that when applied to the data generates an estimate of the parameter (the ATE). An estimator is a random variable (a function of the data) with a distribution
  - an estimator can be biased or unbiased
  - it can have a smaller or lower variance
  - we can use simulations to assess the distribution of the estimator under controlled data-generating distributions
- An estimate is a particular value of the estimator in a given dataset

Parameter versus estimator versus estimate

From the note on GitHub:

We note here that we could also have expressed or "parametrized" the target parameter differently: Using iterated expectations it is straightforward to show both that $\Psi(\mathrm{P}) = \Psi_2(\mu, \pi)$ and $\Psi(\mathrm{P}) = \Psi_3(\mu, f, \pi)$, where

$$\Psi_2(\mu, \pi) := \int \left\{ \frac{a\, y}{\pi(x)} - \frac{(1-a)\, y}{1-\pi(x)} \right\} \mathrm{d}\mu(y, a, x),$$

$$\Psi_3(\mu, f, \pi) := \int \left\{ \frac{a\,(y - f(x,1))}{\pi(x)} + f(x,1) - \frac{(1-a)\,(y - f(x,0))}{1-\pi(x)} - f(x,0) \right\} \mathrm{d}\mu(y, a, x),$$

with $\pi(x) := \mathrm{P}(A = 1 \mid X = x)$ denoting the conditional probability of treatment given covariate status. Hence, using Ψ_2 the nuisance parameters would instead be the treatment mechanism π and the full measure μ, while using Ψ_3 the nuisance parameters would be f, π, and μ. For later reference we define

$$\begin{aligned} &\varphi_1(x; f) := f(1,x) - f(0,x), \quad \varphi_2(y, a, x; \pi) := \frac{a\, y}{\pi(x)} - \frac{(1-a)\, y}{1-\pi(x)}, \quad \text{and} \\ &\varphi_3(y, a, x; f, \pi) := \varphi_2(y, a, x; \pi) + \varphi_1(x; f) - \frac{a\, f(1,x)}{\pi(x)} + \frac{(1-a)\, f(0,x)}{1-\pi(x)}, \end{aligned} \tag{2}$$

such that we can write $\Psi(\mathrm{P}) = \mathrm{P}[\varphi_1(O, f)] = \mathrm{P}[\varphi_2(O, \pi)] = \mathrm{P}[\varphi_3(O, f, \pi)]$.

→ we have different parametrizations of the target parameter (ATE)

Parameter versus estimator versus estimate

Estimation of each can be done via a two-step procedure:

For instance, when estimating the ATE and using the parametrization in (1), we would (1) estimate the conditional outcome $f(x, y) = \mathbb{E}[Y \mid A = a, X = x]$ and the marginal distribution μ_X with estimators $\hat{f}_n$ and $\hat{\mu}_n$, and then (2) plug these into Ψ_1. Estimation of μ_X is straightforward using the empirical measure $\hat{\mathbb{P}}_n$, which gives the estimator

$$\hat{\theta}_n = \Psi_1(\hat{\mathbb{P}}_n, \hat{f}_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{f}_n(1, X_i) - \hat{f}_n(0, X_i) \right\}, \tag{3}$$

where $\hat{f}_n$ is some estimated regression function, for instance obtained by linear regression. Using instead the parametrization given by Ψ_2 would demand estimation of π in step (1), giving the estimator $\Psi_2(\hat{\mathbb{P}}_n, \hat{\pi}_n)$, while using Ψ_3 would demand estimation of both f and π, giving the estimator $\Psi_3(\hat{\mathbb{P}}_n, \hat{f}_n, \hat{\pi}_n)$.

Parameter versus estimator versus estimate

Example of an estimator 1:

- Use a logistic regression model of Y on A, X to obtain an estimator for $f(A, X) = \mathbb{E}[Y \mid A, X]$: $\hat{f}_n(A, X) = \mathrm{expit}(\hat{\beta}_0^Y + \hat{\beta}_A^Y A + X^\top \hat{\beta}_X^Y)$, with covariates X
- Use a logistic regression model of A on X to obtain an estimator for $\pi(X) = \mathbb{E}[A \mid X]$: $\hat{\pi}_n(X) = \mathrm{expit}(\hat{\beta}_0^A + X^\top \hat{\beta}_X^A)$, with covariates X
- Plug in $\hat{f}_n$ and $\hat{\pi}_n$ to obtain the estimator $\hat{\psi}_{3,n} = \Psi_3(\mathbb{P}_n, \hat{f}_n, \hat{\pi}_n)$

Parameter versus estimator versus estimate

Example of an estimator 2:

- Use a logistic regression model of Y on A, X to obtain an estimator for $f(A, X) = \mathbb{E}[Y \mid A, X]$: $\hat{f}_n(A, X) = \mathrm{expit}(\hat{\beta}^Y_{n\,0} + \hat{\beta}^Y_{n\,A} A + X^\top \hat{\beta}^Y_{n\,X})$, with covariates X
- Use a logistic regression model of A on X to obtain an estimator for $\pi(X) = \mathbb{E}[A \mid X]$: $\hat{\pi}_n(X) = \mathrm{expit}(\hat{\beta}^A_{n\,0} + X^\top \hat{\beta}^A_{n\,X})$, with covariates X
- Use a TMLE step where $\hat{f}_n$ is updated using the information from $\hat{\pi}_n$ to obtain the updated estimator $\hat{f}^*_n$ for f and plug this into the g-formula: $\hat{\psi}_{1,n} = \Psi_1(\mathbb{P}_n, \hat{f}^*_n)$

Parameter versus estimator versus estimate

Reporting the final estimate for the ATE:

Based on method X, we estimated an ATE of $0.xx$ *with confidence intervals* $(xx1, xx2)$

We can estimate the variance of an asymptotically linear estimator

Under some conditions (see summary in GitHub note, Section 4), the estimator $\hat{\psi}_{3,n} = \Psi_3(\mathbb{P}_n, \hat{f}_n, \hat{\pi}_n)$ and likewise the TMLE estimator $\Psi_1(\mathbb{P}_n, \hat{f}_n^*)$ is asymptotically linear:

$$\hat{\psi}_n - \psi_0 = \frac{1}{n} \sum_{i=1}^{n} \phi^*(f_0, \pi_0) + o_P(n^{-1/2})$$

(here ϕ^* denotes the efficient influence function)

This implies that

$$\sqrt{n}\left(\hat{\psi}_n - \psi_0\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, P_0 \phi^*(f_0, \pi_0)^2)$$

The asymptotic variance of the estimator can be estimated by $\hat{\sigma}_n^2 = \mathbb{P}_n(\phi^*(\hat{f}_n, \hat{\pi}_n)^2)$ plugging in estimators for f and π into ϕ^*

What is the point with the simulation study?

What is the point with the simulation study?

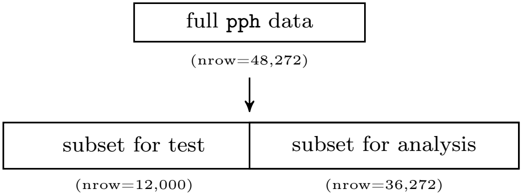
- Note that we would not need semiparametric efficiency theory if data were always generated from a parametric model... but they are usually not
- In our case (the observed data) we are need to control for many covariates — and we would like to have flexible methods to do this
  - A random forest as a good example that does not need any model prespecification
  - A logistic regression, on the other hand requires a prespecified model and we have no prior knowledge to specify it

What is the point with the simulation study?

For exploration, we work with simulated versions of the data

One way to do it:

```
set.seed(12345)
subset.test <- pph[sample(1:nrow(pph), 12000)]
subset.analysis <- pph[!(pph$ID %in% subset.test$ID)]
```



What is the point with the simulation study?

- In the observed data, the forms of $f(a,x) = \mathbb{E}[Y \mid A = a, X = x]$ and $\pi(a \mid x) = P(A = a \mid X = x)$ are *unknown*
- In the simulation study, *we decide what they are*
- We simply generate the data such that:
  - $X \sim \mu_X$ (by sampling from the observed covariates)
  - $\mathbb{E}[A \mid X] = \beta_0^A + X^\top \beta_X^A$
  - $\mathbb{E}[Y \mid X] = \beta_0^Y + \beta_A^Y A + X^\top \beta_X^Y$
- Now we can assess how it affects estimation if we use *correctly* specified logistic regression models versus *misspecified* logistic regression models versus a flexible algorithm such as a random forest

Final data analysis

Final data analysis

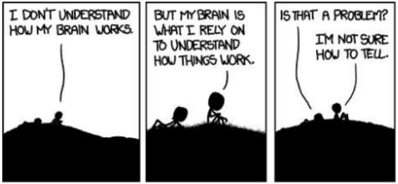
- In the end we want to analyze the *observed* data
- Here we used the remaining part of the data taken out before we started the simulations
- Maybe we want to use the same different estimators as in the simulation study
  - but maybe we know, from the simulation study and the theory, that some or one of the estimators is more robust than the others
  - So, if their conclusions (in terms of ATE) differ in the data analysis, we can explain why that might be

Final data analysis

Reporting the final estimate for the ATE:

> *Based on method X, we estimated an ATE of* $0.xx$ *with confidence intervals* $(xx1, xx2)$

Young Statisticians Denmark



22

TUESDAY, 22 JUNE 2021 FROM 19:00 UTC+02-21:30 UTC+02

Statistics meets neurobiology - talk & quiz

Free · Copenhagen

- A talk by Susanne Ditlevsen about stochastic processes and the brain
- Quiz and social stuff

Link to Facebook event: `https://fb.me/e/X1iezwil`
Otherwise, write Anders: a.munch@sund.ku.dk