Roadmap of Statistical Targeted Learning

Alessandra Meddis
Section of Biostatistics

TMLE for Breakfast
May 22 2023

Why Targeted Learning?

The main goal of Targeted Learning is translating real-world problem into a statistical formulation

- Targeted: the analysis is specifically designed to estimate the quantity of interest which is formulated (defined) based on the question to address

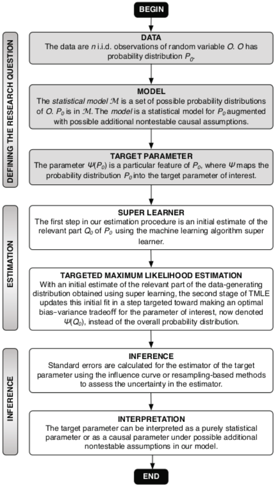
Question (quantity of interest) → Statistical model

Instead of

Statistical model → Question (interpretation of estimated quantity)

We will define everything into a causal framework (counterfactuals) that helps us understanding whether we would be able to address the scientific question

Roadmap of Targeted Learning[1]



DATA
The data are n i.i.d. observations of random variable O. O has probability distribution P_0.

MODEL
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional nontestable causal assumptions.

TARGET PARAMETER
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.

SUPER LEARNER
The first step in our estimation procedure is an initial estimate of the relevant part Q_0 of P_0 using the machine learning algorithm super learner.

TARGETED MAXIMUM LIKELIHOOD ESTIMATION
With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest, now denoted $\Psi(Q_0)$, instead of the overall probability distribution.

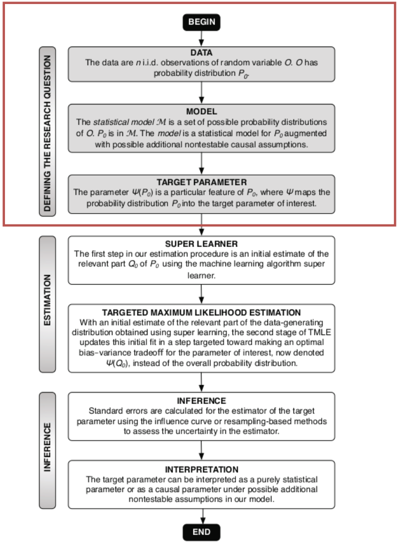
INFERENCE
Standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator.

INTERPRETATION
The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model.

- Scientific question :
  1. Observed data
  2. The Causal model
  3. The Parameter of interest (Target Parameter)
  4. Identifiability
- Statistical part:
  5. Definition of the statistical problem
  6. Estimation

[1] Fig 2.6. Van der Laan, Mark J., and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Vol. 4. New York: Springer, 2011

Roadmap of Targeted Learning



BEGIN

DEFINING THE RESEARCH QUESTION

DATA
The data are n i.i.d. observations of random variable O. O has probability distribution P_0.

MODEL
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional nontestable causal assumptions.

TARGET PARAMETER
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.

ESTIMATION

SUPER LEARNER
The first step in our estimation procedure is an initial estimate of the relevant part Q_0 of P_0 using the machine learning algorithm super learner.

TARGETED MAXIMUM LIKELIHOOD ESTIMATION
With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest, now denoted $\Psi(Q_0)$, instead of the overall probability distribution.

INFERENCE

INFERENCE
Standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator.

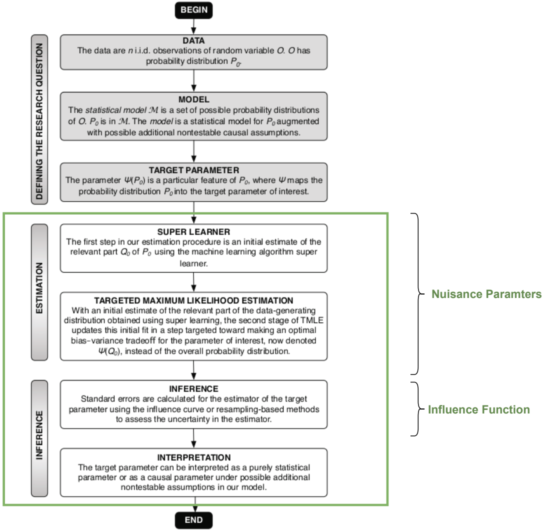
INTERPRETATION
The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model.

END

Registry data in Denmark of patients with diabetes.
We want to compare the risk of cardiovascular disease if treated by GLP1-RA and SGLT2i (a second-line anti-diabetic therapy) or with the standard treatment combination.

Follow-up: 5 years, Competing event: death

Roadmap of Targeted Learning



BEGIN

DEFINING THE RESEARCH QUESTION

DATA
The data are n i.i.d. observations of random variable O. O has probability distribution P_0.

MODEL
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional nontestable causal assumptions.

TARGET PARAMETER
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.

ESTIMATION

SUPER LEARNER
The first step in our estimation procedure is an initial estimate of the relevant part Q_0 of P_0 using the machine learning algorithm super learner.

TARGETED MAXIMUM LIKELIHOOD ESTIMATION
With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest, now denoted $\Psi(Q_0)$, instead of the overall probability distribution.

INFERENCE

INFERENCE
Standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator.

INTERPRETATION
The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model.

END

Nuisance Paramters

Influence Function

Statistical model → Question

Standard approach: Cox model → estimate Hazard Ratios

- Relevant interpretation? (instantaneous risk of event)
- Misspecification problems? (semiparametric approach)
- Biased results? (assumption that treatment is changing at random over time)
- Model selection based on data (goodness of fit)

Question → Statistical model

Targeted Learning → estimate the Target Parameter of interest

- Relevant interpretation? Definition of paramater of interest in a causal framework
- Misspecification problems? Model-free definition of the target parameter
- Biased results? Check for identifiability
- Model selection based on data Machine learning methods + cross validation

Observed data

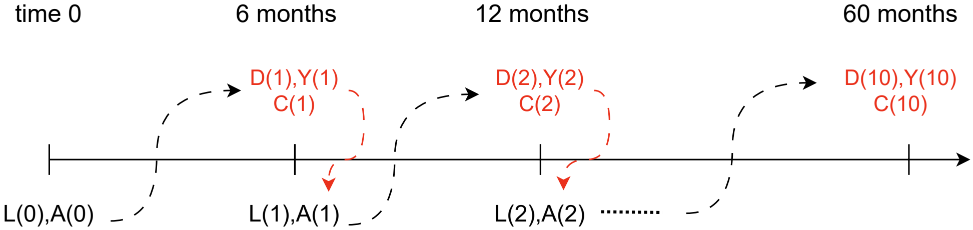
Define which data are available
We discretize the follow-up in k-months length interval.
In each time interval $[k, k+1]$ we observe:

- Covariates: $L(k) \in \mathbb{R}^d$
- Treatment $A(k) \in \{0, 1\}$
- Outcome $Y(k+1) \in \{0, 1\}$
- Competing event $D(k+1) \in \{0, 1\}$
- Censoring status $C(k+1) \in \{0, 1\}$

$$\mathbf{X} = (L(0), A(0), Y(1), D(1), C(1), ..., Y(K), C(K)) \sim P_X$$

Observed data

We discretize the follow-up in 6-months length interval.



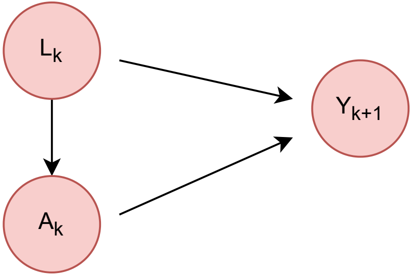
No competing event and No censoring from now on

Discussion for Tomorrow

Causal model

What do we know/assume about the data generation mechanism?

Causal relations among variables?[2]



[2] DAG for time interval $(k, k+1]$

Factorization of Data

$P_X \in \mathcal{P}$ where $\mathcal{P}$ is the set of possible probability distributions that could describe the process by which our observed data have been generated.
We can factorize the joint distribution of the data:

$$P_X(dx) = \prod_{k=1}^{K} P_{Y(k)|\mathcal{F}_{Y(k)}} P_{A(k-1)|\mathcal{F}_{A(k-1)}} P_{L(k-1)|\mathcal{F}_{L(k-1)}}$$

where $\mathcal{F}$ denotes the history/filtration to a given variable, for example $\mathcal{F}_{A(2)} = (L(0), A(0), Y(1), L(1), A(1))$

Factorization of Data

$P_X \in \mathcal{P}$ where $\mathcal{P}$ is the set of possible probability distributions that could describe the process by which our observed data have been generated.
We can factorize the joint distribution of the data:

$$P_X(dx) = \prod_{k=1}^{K} \underbrace{P_{Y(k)|\mathcal{F}_{Y(k)}}}_{Q_{Y(k)}(y|a,l)} P_{A(k-1)|\mathcal{F}_{A(k-1)}} dP_{L(k-1)|\mathcal{F}_{L(k-1)}}$$

where $\mathcal{F}$ denotes the history/filtration to a given variable, for example $\mathcal{F}_{A(2)} = (L(0), A(0), Y(1), L(1), A(1))$

Factorization of Data

$P_X \in \mathcal{P}$ where $\mathcal{P}$ is the set of possible probability distributions that could describe the process by which our observed data have been generated.

We can factorize the joint distribution of the data:

$$P_X(dx) = \prod_{k=1}^{K} P_{Y(k)|\mathcal{F}_{Y(k)}} \underbrace{P_{A(k-1)|\mathcal{F}_{A(k-1)}}}_{G_{A(k)}(a|L=l)} P_{L(k-1)|\mathcal{F}_{L(k-1)}}$$

where $\mathcal{F}$ denotes the history/filtration to a given variable, for example $\mathcal{F}_{A(2)} = (L(0), A(0), Y(1), L(1), A(1))$

Causal framework:

Causal Inference is related to counterfactuals (potential outcomes).
We introduce counterfactuals for one intervention with two levels (0,1 / treated, non-treated) without the longitudinal aspect

Each individual has two potential outcomes:
$Y^{a=0}$: Outcome if allocated to treatment $a = 0$
$Y^{a=1}$: Outcome if allocated to treatment $a = 1$

Subjects receive at most one of the treatment, so at least one will not be observed → counterfactual.

We have interpretation for: outcome if individuals had received the intervention of interest.

Parameter of interest

What do we want to learn from the data?
Define the causal question formulating the scientific problem into counterfactuals from an hypothetical experiment

Define the target parameter: $\Psi : \tilde{P} \rightarrow R$, where $\tilde{P} \in \tilde{\mathcal{P}}$ is the probability distribution for the hypothetical population of counterfactuals

$$(L(0), Y^1(1), Y^0(1), L^1(1), L^0(1), ..., Y^1(10), Y^0(10)) \sim \tilde{P}$$

Parameter of interest

In our example

- Scientific problem: Evaluate whether exposure to the dual second-line treatment with d GLP1-RA and SGLT2i (A=1) increase the risk of cardiovascular disease after 5 years in diabetic patients
- Hypothetical experiment: as a function of counterfactuals
  1. Assigning A=1 continuously to the whole population and observing the outcome after 5 years $Y^1(10)$
  2. Assigning A=0 continuously to the whole population and observing the outcome after 5 years $Y^0(10)$

→ Target Parameter: 5-years Risk difference under the two treatment assignments

$$\Psi(\tilde{P}) = ATE = \tilde{P}(Y^1(10) = 1) - \tilde{P}(Y^0(10) = 1)$$

Identifiability

Can we estimate the target parameter from the observed data?

We cannot observe both counterfactual outcomes for each individual because we can assign either exposure or no exposure to one individual.
$\rightarrow$ We cannot directly estimate counterfactual outcomes but we need some assumptions under which the target parameter is identifiable and so it may be estimated from the observed data:
Define with a^* the treatment intervention

- Exchangeability: $Y^{a^*(k)} \perp A(k)_{|\mathcal{F}_{A(k)}} \forall k$
- Consistency: $Y^{a^*(k)} = Y$ if $A(k) = a^*(k)\ \forall k$
- Positivity: $0 < P_{A(k)|\mathcal{F}_{A(k)}}(A(k) = a^*(k)) < 1\ \forall k$

Identifiability

Can we estimate the target parameter from the observed data?

- Exchangeability: $Y^{a^*(k)} \perp A(k)_{|\mathcal{F}_{A(k)}} \forall k$
- Consistency: $Y^{a^*(k)} = Y$ if $A(k) = a^*(k)$ $\forall k$
- Positivity: $0 < P_{A(k)|\mathcal{F}_{A(k)}}(A(k) = a^*(k)) < 1$ $\forall k$

When these assumptions are verified we can write the parameter as a function of the observed data

$$\tilde{P}(Y^1(10) = 1) = P(Y(10) = 1|\overline{A}(10) = 1)$$

where $\overline{A}(10) = a$ is equal to assign $A(k) = a$ $\forall k$

Definition of the statistical problem

Nuisance Parameters: what do we need as support for the estimation of the target parameter?
Under some calculations $\tilde{P}(Y(10) = 1)$ can be rewritten as a sequence of nested expectation/integrals of the Q and the G functions at different times.

$$\tilde{P}(Y^1(10) = 1) = \int \dots \int Q_{Y(10)}(1|a(k-1), l(k-1))\times \prod_{k=1}^{9} Q_{Y(k)}(dy(k)|1, l(k)) G^*_{A(k)}(da(k)|l(k), a(k-1)) Q_{L(k)}(dl(k)|1, l(k-1))$$

with $G^*_{A(k)}$ denotes the counterfactual distribution for $A(k) = a^*(k)$

6. Statistical Estimation

For the ATE the nuisance parameters are:

- Propensity score model (G-function)
- Outcome model (Q-function)

How can we estimate the nuisance parameters?

- glm (logistic regression)
- Machine learning method
- Superlearner

Statistical estimation

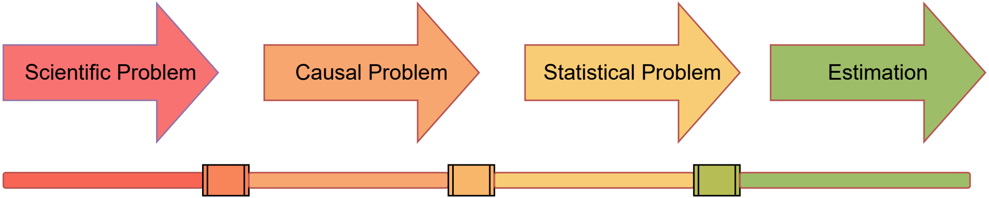
Uncertainty of $\hat{\Psi}$: How do we construct Confidence Interval?

$$\hat{\Psi} - \Psi = \frac{1}{n} \sum_{i=1}^{n} IC(X_i, \nu) + o_p(n^{-1/2})$$

where $IC(X_i, \nu)$ is the influence function that depends on the observed data and the nuisance parameters ν

Sum up

Targeted Learning : from the research question to the estimation of the parameter of interest (Target Parameter)



Scientific Problem

- Question to address
- Observed Data

Causal Problem

- Causal Model
- Target Parameter

Statistical Problem

- Statistical Parameter
- Statistical Model

Estimation

- Nuisance Parameter
- Targeting Algorithm