Day 1, Lecture 4

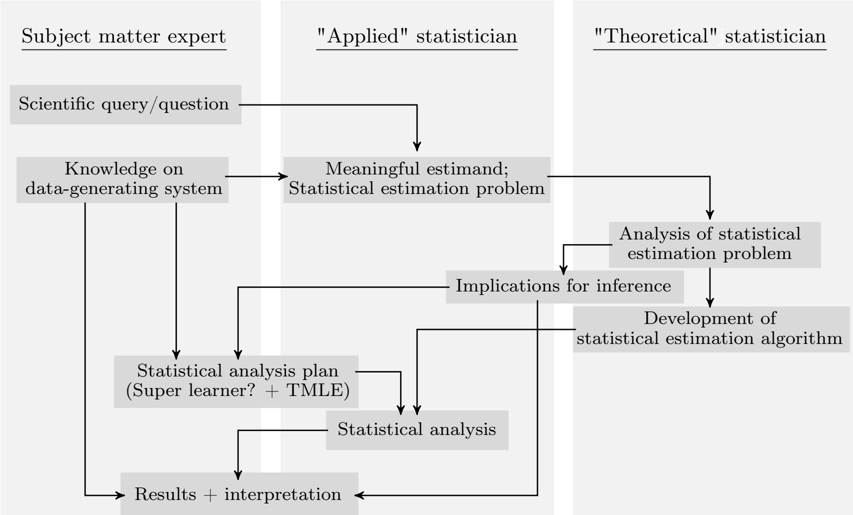
Targeted nonparametric inference

Targeted nonparametric inference

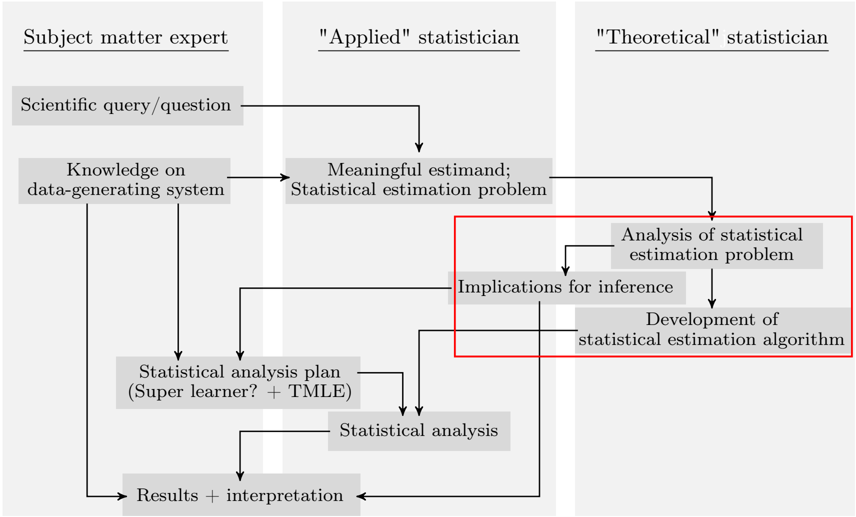
In this lecture, our goal is to:

1. Relate fundamental concepts of nonparametric efficiency theory to the construction of asymptotically linear estimators.
2. Explain the particular role of the efficient influence curve in guiding the development of estimators, including the significance of second order remainders in achieving asymptotic linearity.

Targeted nonparametric inference


Subject matter expert
"Applied" statistician
"Theoretical" statistician
Scientific query/question
Knowledge on data-generating system
Meaningful estimand; Statistical estimation problem
Analysis of statistical estimation problem
Implications for inference
Development of statistical estimation algorithm
Statistical analysis plan (Super learner? + TMLE)
Statistical analysis
Results + interpretation

Targeted nonparametric inference


Subject matter expert
"Applied" statistician
"Theoretical" statistician
Scientific query/question
Knowledge on data-generating system
Meaningful estimand; Statistical estimation problem
Analysis of statistical estimation problem
Implications for inference
Development of statistical estimation algorithm
Statistical analysis plan (Super learner? + TMLE)
Statistical analysis
Results + interpretation

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

- asympotically linear estimation
- efficient influence curve
- second-order remainders

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

- asympotically linear estimation
- efficient influence curve
- second-order remainders

While TMLE can be applied without knowing about these concepts ...

... we cannot understand the purpose of TMLE without knowing a bit about the efficient influence curve.

... to understand what conditions are required for TMLE inference, we need to understand the second-order remainder.

Targeted nonparametric inference

Important notation:[1]

- ▷ For a function $h : \mathcal{O} \to \mathbb{R}$ and distribution P

$$Ph = \mathbb{E}_P[h(O)] = \int h dP = \int_{\mathcal{O}} h(o) dP(o)$$

where $\mathcal{O} = \mathbb{R}^d \times \{0,1\} \times \{0,1\}$ is the sample space of $O = (X, A, Y)$.

- ▷ For the empirical measure $\mathbb{P}_n$ of the sample $O_1, \ldots, O_n$:

$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} h(O_i);$$

note: the right-hand-side is really just the empirical average.

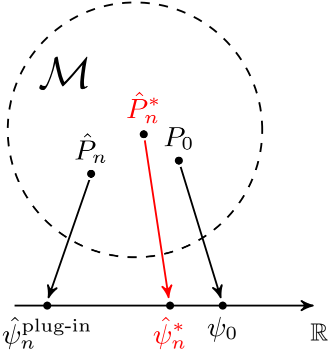
- ▷ $X_n = o_P(1)$ means that $X_n \overset{P}{\to} 0$; $X_n = o_P(n^{-1/2})$ means that $n^{1/2} X_n \overset{P}{\to} 0$.

[1] van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

Targeted nonparametric inference

Recap statistical "roadmap":

1. Data is a random variable O with a probability distribution P_0
2. P_0 belongs to a statistical model $\mathcal{M}$
3. Our target is a parameter $\Psi : \mathcal{M} \to \mathbb{R}$
4. Construct estimator $\hat{P}_n$ for (relevant part of) P_0 and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$
5. Quantify uncertainty for the estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$



Asymptotic linearity

A very desirable property —

[2] recall: $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

The empirical measure $\mathbb{P}_n$ of the sample $O_1, \dots, O_n$:

$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} h(O_i).$$

A very desirable property —

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if [2]

$$\sqrt{n}\big(\hat{\psi}_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

[2] recall: $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

A very desirable property —

The empirical measure $\mathbb{P}_n$ of the sample $O_1, \ldots, O_n$:

$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} h(O_i).$$

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if [2]

$$\sqrt{n}\left(\hat{\psi}_n - \psi_0\right) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

Then CLT + Slutsky implies:

$$\hat{\psi}_n \overset{as}{\sim} N(\psi_0, \mathrm{Var}(\phi(P_0))/n).$$

The estimator behaves asymptotically as an average of the influence function.

[2] recall: $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

I.e., as n gets big, the difference between our estimate and the truth behaves as if it were an average of iid variables, which, by the CLT, converges to a (stable) normal distribution.

Of course, the approximation is not perfect. At any finite n, there is some (random) "second order" remainder,

$$\hat{\psi}_n - \psi_0 - \sqrt{n}\phi^*(P_0).$$

We get back to this later in the course.

The following aims to provide a bit of intuition. →

Asymptotic linearity

An estimator for the mean $\psi_0 = \mathbb{E}_{P_0}[X]$:

$$\hat{\psi}_{n,0} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then

$$\sqrt{n}(\hat{\psi}_{n,0} - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} = \sqrt{n}\mathbb{P}_n\phi(P_0)$$

$\hat{\psi}_{n,0}$ is linear and thus asymptotically linear.

Asymptotic linearity

An estimator for the mean $\psi_0 = \mathbb{E}_{P_0}[X]$:

$$\hat{\psi}_{n,1} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}$$

Then

$$\sqrt{n}(\hat{\psi}_{n,1} - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{\frac{1}{\sqrt{n}}}_{=o(1)}$$

$\hat{\psi}_{n,1}$ is asymptotically linear.

Asymptotic linearity

An estimator for the mean $\psi_0 = \mathbb{E}_{P_0}[X]$:

$$\hat{\psi}_{n,2} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n^{1/2+0.1}}$$

Then

$$\sqrt{n}(\hat{\psi}_{n,2} - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n^{1/2+0.1}} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{\frac{1}{n^{0.1}}}_{=o(1)}$$

$\hat{\psi}_{n,2}$ is asymptotically linear.

Asymptotic linearity

An estimator for the mean $\psi_0 = \mathbb{E}_{P_0}[X]$:

$$\hat{\psi}_{n,3} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n^{1/2-0.1}}$$

Then

$$\sqrt{n}(\hat{\psi}_{n,3} - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n^{1/2-0.1}} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{n^{0.1}}_{\to\infty}$$

$\hat{\psi}_{n,3}$ is **not** asymptotically linear.

Asymptotic linearity

An estimator $\hat{\psi}_n$ has rate of convergence $r_n \to \infty$ if [3]

$$r_n(\hat{\psi}_n - \psi_0) = O_P(1), \quad \text{i.e.,} \quad \hat{\psi}_n - \psi_0 = O_P(1/r_n).$$

The convergence rate r_n tells us how fast $\hat{\psi}_n$ centers around ψ_0, with the difference $\hat{\psi}_n - \psi_0$ behaving like $1/r_n$.

- One wants negligible bias such as to obtain reliable confidence intervals for ψ_0.
- The bias of an asymptotically linear estimator converges to zero at a rate faster than $1/\sqrt{n}$.

Data-adaptive machine learning estimators rarely achieve this rate.

[3] recall: $O_P(1)$ denotes a sequence which is bounded in probability.

Asymptotic linearity

$$\sqrt{n}\hat{\psi}_{n,1} = \sqrt{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n}}_{\to 0}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,1} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,2} = \sqrt{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2+0.1}}}_{\to 0}, \quad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,3} = \sqrt{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2-0.1}}}_{\to \infty}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) \overset{P}{\to} \infty.$$

Asymptotic linearity

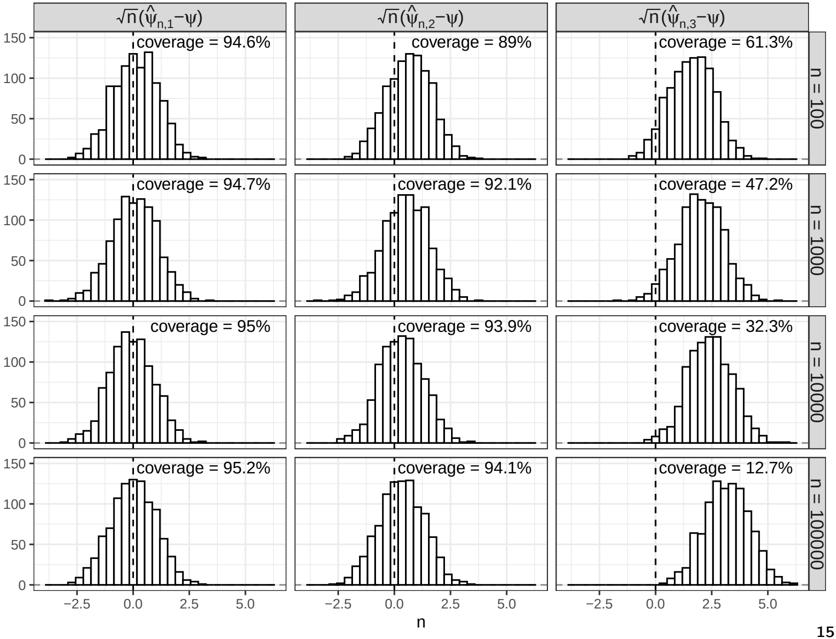
$$\sqrt{n}\hat{\psi}_{n,1} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n}}_{\to 0}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,1} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,2} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2+0.1}}}_{\to 0}, \quad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,3} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2-0.1}}}_{\to \infty}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) \overset{P}{\to} \infty.$$

[The remainder term that determines the asymptotic bias the estimator].

Asymptotic linearity



Estimator expansion and the efficient influence curve

Repetition — our goal is $\sqrt{n}$-consistency and asymptotic linearity.

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if

$$\sqrt{n}\big(\hat{\psi}_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

Then CLT + Slutsky implies:

$$\hat{\psi}_n \overset{as}{\sim} N(\psi_0, \mathrm{Var}(\phi(P_0))/n).$$

The estimator behaves asymptotically as an average of the influence function.

Estimator expansion and the efficient influence curve

A key component in constructing a $\sqrt{n}$-consistent and asymptotically linear estimator, *even when using machine learning estimation*, is the so-called the efficient influence curve.[4]

(Don't) be confused. We both talk about:

- the influence function for an *estimator*.
- the efficient influence function (curve) for the *estimation problem*.

[4] also known as the efficient influence function, the pathwise derivative, the Neyman orthogonal score, the canonical gradient.

Estimator expansion and the efficient influence curve

The von Mises expansion:

A sufficiently smooth $\Psi : \mathcal{M} \to \mathbb{R}$ admits a certain distributional Taylor expansion:

$$\Psi(P) - \Psi(P') = \int \phi(P)(o) d(P - P')(o) + R_2(P, P'), \qquad (1)$$

for distributions $P, P' \in \mathcal{M}$ and a function ϕ satisfying $P\phi(P) = 0$ (mean zero) and $P\phi(P)^2 < \infty$ (finite variance).

We will use (1) as a starting-point to start analyzing:

$$\hat{\psi}_n - \psi_0 = \Psi(\hat{P}_n) - \Psi(P_0).$$

Estimator expansion and the efficient influence curve

- when the model $\mathcal{M}$ is assumed proper nonparametric, there exists *one* function $\phi(P)$ fulfilling (1). This is called the efficient influence curve; we also denote it $\phi^*(P)$.

Estimator expansion and the efficient influence curve

- when the model $\mathcal{M}$ is assumed proper nonparametric, there exists *one* function $\phi(P)$ fulfilling (1). This is called the efficient influence curve; we also denote it $\phi^*(P)$.
    - this may be confusing here, but in restricted (semi)parametric models, multiple ϕ's can satisfy (1).

Estimator expansion and the efficient influence curve

- when the model $\mathcal{M}$ is assumed proper nonparametric, there exists *one* function $\phi(P)$ fulfilling (1). This is called the efficient influence curve; we also denote it $\phi^*(P)$.
  - this may be confusing here, but in restricted (semi)parametric models, multiple ϕ's can satisfy (1).
  - for these situations we by the way have that $P_0\phi(P_0)^2 \geq P_0\phi^*(P_0)^2$.
    (which is why it is called the efficient influence curve).

Estimator expansion and the efficient influence curve

- when the model $\mathcal{M}$ is assumed proper nonparametric, there exists *one* function $\phi(P)$ fulfilling (1). This is called the efficient influence curve; we also denote it $\phi^*(P)$.
  - this may be confusing here, but in restricted (semi)parametric models, multiple ϕ's can satisfy (1).
  - for these situations we by the way have that $P_0\phi(P_0)^2 \geq P_0\phi^*(P_0)^2$.
    (which is why it is called the efficient influence curve).
- the efficient influence curve in nonparametric models indicates how to construct asymptotically linear (and efficient) estimators . . .

Estimator expansion and the efficient influence curve

- when the model $\mathcal{M}$ is assumed proper nonparametric, there exists *one* function $\phi(P)$ fulfilling (1). This is called the efficient influence curve; we also denote it $\phi^*(P)$.
  - this may be confusing here, but in restricted (semi)parametric models, multiple ϕ's can satisfy (1).
  - for these situations we by the way have that $P_0\phi(P_0)^2 \geq P_0\phi^*(P_0)^2$.

    (which is why it is called the efficient influence curve).
- the efficient influence curve in nonparametric models indicates how to construct asymptotically linear (and efficient) estimators . . . the goal is to construct an asymptotically linear estimator which has influence function equal to the efficient influence curve.

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\Psi(\hat{P}_n) - \Psi(P_0) = (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0)$$

$$(*1)$$

$$(*2)$$

$$(*3)$$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0)\\ &= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n)\end{aligned}$$

$$(*1)$$

$$(*2)$$

$$(*3)$$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0)\\ &= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n)\\ &\quad \pm \mathbb{P}_n\phi^*(\hat{P}_n)\\ &\quad \pm (\mathbb{P}_n - P_0)\phi^*(P_0)\end{aligned}$$

$(*1)$

$(*2)$

$(*3)$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0) \\
&= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n) \\
&\qquad \pm\ \mathbb{P}_n\phi^*(\hat{P}_n) \\
&\qquad \pm\ (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&\qquad +\ (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n) - \phi^*(P_0)\big) && (*1) \\
&\qquad +\ R_2(\hat{P}_n, P_0) && (*2) \\
&\qquad -\ \mathbb{P}_n\phi^*(\hat{P}_n) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0) \\
&= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n) \\
&\qquad \pm\ \mathbb{P}_n\phi^*(\hat{P}_n) \\
&\qquad \pm\ (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) \\
&\qquad +\ (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n) - \phi^*(P_0)\big) && (*1) \\
&\qquad +\ R_2(\hat{P}_n, P_0) && (*2) \\
&\qquad -\ \mathbb{P}_n\phi^*(\hat{P}_n) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0) \\
&= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n) \\
&\qquad \pm \mathbb{P}_n\phi^*(\hat{P}_n) \\
&\qquad \pm (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n) - \Psi(P_0) &= (\hat{P}_n - P_0)\phi^*(\hat{P}_n) + R_2(\hat{P}_n, P_0) \\
&= -P_0\phi^*(\hat{P}_n) + R_2(P_0, \hat{P}_n) \\
&\qquad \pm \mathbb{P}_n\phi^*(\hat{P}_n) \\
&\qquad \pm (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n) && (*3)
\end{aligned}$$

i.e., need $(*1)$–$(*3)$ to be $o_P(n^{-1/2})$.

Estimator expansion

An estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear if,

$$\sqrt{n}\big(\Psi(\hat{P}_n) - \Psi(P_0)\big) = \sqrt{n}\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

$$\begin{aligned}\Psi(\hat{P}_n) - \Psi(P_0) = \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) & \\ + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n) - \phi^*(P_0)\big) & \qquad (*1) \\ + R_2(\hat{P}_n, P_0) & \qquad (*2) \\ - \mathbb{P}_n\phi^*(\hat{P}_n) & \qquad (*3)\end{aligned}$$

- $(*1)$ is an empirical process term.
- $(*2)$ second-order bias term.
- $(*3)$ is called the efficient influence curve equation.

Estimator expansion

... about the empirical process term $(*1)$:

1. can be handled by empirical process theory, if $(\phi^*(P) : P \in \mathcal{M})$ is assumed Donsker.[5]
2. otherwise can handled by extra sample splitting.[6, 7]

[5] Lemma 19.24 of van der Vaart, A. W. (2000): Asymptotic statistics yields then that $(\mathbb{P}_n - P_0)(\phi^*(\hat{P}_n) - \phi^*(P_0)) = o_P(n^{-1/2})$.

[6] Zheng, W., & van der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation.

[7] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Estimator expansion

Side note: Usually, we will assume the Donsker class condition.

- this is a way of nonparametrically characterizing the complexity of nuisance parameters.
- classes of functions that are Donsker: Indicator functions, bounded monotone functions, Lipschitz parametric functions, smooth functions, ...

Donsker classes also include traditional parametric functions.

We will not discuss this further. For a nice intro see Sections 4.2 and 4.3 of Kennedy, E. H. (2016): Semiparametric theory and empirical processes in causal inference.

Estimator expansion

This is basically what is needed to understand the construction of TMLE.

Estimator expansion

This is basically what is needed to understand the construction of TMLE.

Conditions (asymptotic linearity and efficiency)

(C1) Solve the efficient influence curve equation: $\mathbb{P}_n \phi^*(\hat{P}_n) = o_P(n^{-1/2})$.

(C2) Remainder $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$.

(C3) Donsker class conditions for $\{\phi^*(P) : P \in \mathcal{M}\}$.

Then: $\Psi(\hat{P}_n) \overset{as}{\sim} N(\Psi(P_0), P_0\phi^*(P_0)^2/n)$.

Construction of estimators

$$\Psi(\hat{P}_n) - \Psi(P_0) = \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2})$$
$$+ R(\hat{P}_n, P_0)$$
$$- \mathbb{P}_n\phi^*(\hat{P}_n)$$

For **a given target parameter** $\Psi : \mathcal{M} \to \mathbb{R}$, we need to

1. Derive the efficient influence curve, so that we can solve the efficient influence curve equation.
2. Analyze the remainder $R(P, P_0) := \Psi(P) - \Psi(P_0) + P_0\phi^*(P)$.

Repetition: These are solely properties of the estimation problem, but also tell us how to construct estimators such as TMLE.

Example: ATE estimation

Analysis of a concrete estimation problem

EXAMPLE: Average treatment effect (ATE)

Observed data $O = (X, A, Y) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\} = \mathcal{O}$

- $X \in \mathbb{R}^d$ are covariates
- $A \in \{0,1\}$ is a binary exposure variable (treatment decision)
- $Y \in \{0,1\}$ is a binary outcome variable

$O \sim P_0$ where P_0 assumed to belong to nonparametric model $\mathcal{M}$.

We are interested in estimating the ATE:

$$\Psi(P) = \mathbb{E}_P\big[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]\big].$$

Analysis of a concrete estimation problem

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\begin{aligned}\phi^*(P)(O) &= \tilde{\phi}^*(f,\pi)(O)\\ &= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)\end{aligned}$$

Analysis of a concrete estimation problem

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\begin{aligned}\phi^*(P)(O) &= \tilde{\phi}^*(f,\pi)(O)\\ &= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)\end{aligned}$$

2. The remainder:

$$\begin{aligned}R(P,P_0) &= \tilde{R}(f,\pi,f_0,\pi_0)\\ &= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big) d\mu_{0,X}(x)\end{aligned}$$

Analysis of a concrete estimation problem

$$f(A,X) = \mathbb{E}_P[Y \mid A,X],\ \pi(A \mid X) = P(A = a \mid X)$$
$$f_0(A,X) = \mathbb{E}_{P_0}[Y \mid A,X],\ \pi_0(A \mid X) = P_0(A = a \mid X)$$

$$R(P,P_0) := \Psi(P) - \Psi(P_0) + P_0\phi^*(P).$$

2. Deriving the remainder for the ATE:

$$\begin{aligned}
R(P,P_0) &= \underbrace{\mathbb{E}_P[f(1,X) - f(0,X)]}_{=\Psi(P)} - \underbrace{\mathbb{E}_{P_0}[f_0(1,X) - f_0(0,X)]}_{=\Psi(P_0)} \\
&+ \mathbb{E}_{P_0}\left[\left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)(Y - f(A,X))\right] \\
&+ \mathbb{E}_{P_0}[f(1,X) - f(0,X)] - \Psi(P) \\
&\overset{*}{=} \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\left(\frac{\pi_0(a \mid x)}{\pi(a \mid x)} - 1\right)(f_0(a,x) - f(a,x))\, d\mu_{0,X}(x) \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}(f_0(a,x) - f(a,x))\, d\mu_{0,X}(x)
\end{aligned}$$

the equality marked by $*$ is detailed on the next slide.

Analysis of a concrete estimation problem

We used that:

$$
\begin{aligned}
&\mathbb{E}_{P_0}\left[\left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big)\right] \\
&\quad = \mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(Y - f(A,X)\big)\right] \\
&\quad = \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(Y - f(A,X)\big)\,\middle|\, A, X\right]\right] \\
&\quad = \mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(f_0(A,X) - f(A,X)\big)\right] \\
&\quad = \int_{\mathbb{R}^d}\sum_{a=0,1}\frac{2a-1}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big)\pi_0(a \mid x)\,d\mu_{0,X}(x) \\
&\quad = \int_{\mathbb{R}^d}\sum_{a=0,1}(2a-1)\frac{\pi_0(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big)\,d\mu_{0,X}(x)
\end{aligned}
$$

Analysis of a concrete estimation problem

The remainder determines the asymptotic bias.

For the ATE, the remainder has a really nice structure.

$$
\begin{aligned}
R(P, P_0) &= \tilde{R}(f, \pi, f_0, \pi_0) \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1) \frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)} \big(f_0(a, x) - f(a, x) \big) d\mu_{0,X}(x)
\end{aligned}
$$

A "double robust" structure, which has some important implications.

Analysis of a concrete estimation problem

$$
\begin{aligned}
&|R(P, P_0)| = |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x)
\end{aligned}
$$

Analysis of a concrete estimation problem

$$
\begin{aligned}
|R(P, P_0)| &= |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x) \\
&\overset{*}{\leq} \sum_{a=0,1} \frac{1}{\pi(a \mid x)} \sqrt{\int_{\mathbb{R}^d} \left\{\pi_0(a \mid x) - \pi(a \mid x)\right\}^2 d\mu_{0,X}(x)} \\
&\qquad \times \sqrt{\int_{\mathbb{R}^d} \left\{f_0(a, x) - f(a, x)\right\}^2 d\mu_{0,X}(x)}
\end{aligned}
$$

$*$ uses Cauchy-Schwarz.

Analysis of a concrete estimation problem

$$
\begin{aligned}
|R(P, P_0)| &= |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x) \\
&\overset{*}{\leq} \sum_{a=0,1} \frac{1}{\pi(a \mid x)} \sqrt{\int_{\mathbb{R}^d} \left\{\pi_0(a \mid x) - \pi(a \mid x)\right\}^2 d\mu_{0,X}(x)} \\
&\qquad \times \sqrt{\int_{\mathbb{R}^d} \left\{f_0(a, x) - f(a, x)\right\}^2 d\mu_{0,X}(x)}
\end{aligned}
$$

Thus, if $\pi(a \mid X) > \delta > 0$ a.s., then:

$$
|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \left\| \pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot) \right\|_{\mu_0} \left\| f_0(a \mid \cdot) - \hat{f}_n(a \mid \cdot) \right\|_{\mu_0}
$$

$*$ uses Cauchy-Schwarz.

Analysis of a concrete estimation problem

What does this imply for estimation?

Double robustness in consistency

$$|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \underbrace{\|\pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot)\|_{\mu_0}}_{o_P(1), \text{ or}} \underbrace{\|f_0(a \mid \cdot) - \hat{f}_n^*(a \mid \cdot)\|_{\mu_0}}_{o_P(1)}$$

then $\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = o_P(1)$.

Asymptotic linearity (easier to establish due to double robust structure)

$$|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \underbrace{\|\pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot)\|_{\mu_0}}_{=o_P(n^{-1/4})} \underbrace{\|f_0(a \mid \cdot) - \hat{f}_n^*(a \mid \cdot)\|_{\mu_0}}_{=o_P(n^{-1/4})}$$

i.e., $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = o_P(n^{-1/2})$.

I.e., bias is converging at fast enough rate for reliable confidence intervals.

Analysis of a concrete estimation problem

Side note: Showing the double robustness in consistency ...

Say we have estimators $(\hat{f}_n^*, \hat{\pi}_n)$;

- converging to (f, π)
- solving the efficient influence curve equation.

Per definition, $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = \tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) + P_0\tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n)$.

I.e., $\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = -P_0\tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n) + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)$

$$= (\mathbb{P}_n - P_0)\phi^*(\hat{f}_n^*, \hat{\pi}_n) + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0).$$

The first term is an empirical process term which equals $(\mathbb{P}_n - P_0)\tilde{\phi}^*(f, \pi)$ plus an $o_P(n^{-1/2})$-term.

This then gives

$$\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = \underbrace{(\mathbb{P}_n - P_0)\tilde{\phi}^*(f, \pi)}_{\text{LLN applies}} + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) + o_P(n^{-1/2})$$

which yields that $\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = o_P(1)$ if $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = o_P(1)$.

Analysis of a concrete estimation problem

Double robustness in consistency

- consistent estimation of target parameter requires only that either f or π is consistently estimated.

can be important, but even more important is:

"Rate double robustness"

- convergence rate requirements for estimators f and π are seriously weakened;
- asymptotic linearity is achieved when these are estimated (consistently) at rate at least $n^{-1/4}$.[8]

[8] and this is actually possible with certain machine learning algorithms; and if these are combined in super learner there is something called the oracle property, which says that the super learner achieves the rate of convergence of the *best* estimator in its library.

Analysis of a concrete estimation problem

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\begin{aligned}\phi^*(P)(O) &= \tilde{\phi}^*(f,\pi)(O) \\ &= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)\end{aligned}$$

2. The remainder:

$$\begin{aligned}R(P,P_0) &= \tilde{R}(f,\pi,f_0,\pi_0) \\ &= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big) d\mu_{0,X}(x)\end{aligned}$$

Deriving these is done once for a given target parameter $\Psi : \mathcal{M} \to \mathbb{R}$.

TMLE as an estimation procedure

TMLE is a two-step procedure:

Step 1 Construct initial estimator $\hat{P}_n$ for P such that $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$.

Step 2 Update the estimator $\hat{P}_n \mapsto \hat{P}_n^*$ such that $\hat{P}_n^*$ solves the efficient influence curve equation.

TMLE and the estimating equation (EE) estimator both solve the efficient influence curve equation . . . that is why they share the same asymptotic (large-sample) properties.

Aside: deriving the efficient influence curve

We will talk more later about deriving efficient influence curves, but just note for now that it is very much tied to the notion of derivatives along smooth parametric submodels.

Aside: deriving the efficient influence curve

We will talk more later about deriving efficient influence curves, but just note for now that it is very much tied to the notion of derivatives along smooth parametric submodels.

The von Mises expansion (1) implies a related notion of smoothness called pathwise differentiability, i.e.,

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \Psi(P_\varepsilon) = \int \phi(P)(o) b(o) dP(o), \tag{2}$$

for every smooth submodel $\{P_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$, where P_ε has density density p_ε, and for which $P_{\varepsilon=0} = P_0$.

Aside: deriving the efficient influence curve

We will talk more later about deriving efficient influence curves, but just note for now that it is very much tied to the notion of derivatives along smooth parametric submodels.

The von Mises expansion (1) implies a related notion of smoothness called pathwise differentiability, i.e.,

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \Psi(P_\varepsilon) = \int \phi(P)(o) b(o) dP(o), \tag{2}$$

for every smooth submodel $\{P_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$, where P_ε has density density p_ε, and for which $P_{\varepsilon=0} = P_0$.

Equation (2) can be used to derive the efficient influence curve.

Aside: deriving the efficient influence curve

For example one evaluates the pathwise derivative along submodels defined for a mean-zero function $h : \mathcal{O} \to \mathbb{R}$ as:

$$p_\varepsilon(o) = p(o)(1 + \varepsilon h(o)).$$

and solves the integral equation on the right hand side of (2).

Note that this submodel has score function $\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \log p_\varepsilon(o) = h(o)$.

(And there are many other useful tricks in this process).

Aside: deriving the efficient influence curve

Another common strategy is to assume the data are discrete and then compute the Gateaux derivative:

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \Psi((1-\varepsilon)p(o)+\varepsilon\delta_{o'}),$$

which equals the influence curve $\phi(P)(o')$ directly.

Note that this really corresponds to computing the pathwise derivative along the particular submodel of the form $(1-\varepsilon)p(o)+\varepsilon\delta_{o'}$ for which the right hand side of (2) is $\phi(P)(o')$.

Practical 2: Continued explorations based on simulated data

In this exercise we continue the simulation setting of Practical 1, now to explore —

1. Inference for estimators based on the efficient influence curve;
2. Variance estimation and coverage of confidence intervals;
3. Small-sample properties, particularly under positivity violations.

This is described in detail in: **day1-practical2.pdf**.

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;
- there may be small-sample differences in performance (often argued in favor of TMLE);

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;
- there may be small-sample differences in performance (often argued in favor of TMLE);
- otherwise, the differences are not so important for the ATE estimation problem. BUT, for other problems (e.g., in survival analysis), the substitution property of TMLE may be crucial.

Summary & takeaways

"Classical" causal inference estimators:

- consistency of g-formula estimators rely on correct specification of the outcome regression $f(a,x) = \mathbb{E}[Y \mid A = a, X = x]$
- consistency of ipw estimators rely on correct specification of the propensity score $\pi(a \mid x) = P(A = a \mid X = x)$
- both types of estimators work poorly with machine learning (cross-validation does not help)

TMLE (and other estimators based on the efficient influence curve):

- built-in bias correction
- double robustness in consistency
- inference based on the efficient influence curve
- even when incorporating machine learning (under conditions!!)

Summary & takeaways

- In the end, we want asymptotic linearity and regularity for our estimator for the low-dimensional target parameter.
- Asymptotic linearity allows normal approximations and confidence intervals.
- The rate of convergence of an estimator $\hat{P}_n$ tells us how fast $\hat{P}_n$ centers around P_0.
- The rate $n^{1/2}$ is typically the fastest we can expect and is referred to as the "parametric rate" of convergence. This is the rate required for asymptotically linear estimation.
- Nuisance estimators do not need parametric rates, but the estimator for the target parameter does.

Summary & takeaways

- Directly plugging in an estimator $\hat{P}_n$ into a functional Ψ will typically lead to an estimator $\Psi(\hat{P}_n)$ that converges to $\psi_0 = \Psi(P_0)$ at the same rate as $\hat{P}_n$ converges to P_0.
- Flexible, data-adaptive estimators (such as kernel-based regression and random forests) rely on fewer assumption than classical parametric models, but they converge at a slower rate.
- If P_0 is estimated using, e.g., a random forest, the asymptotic distribution of $\hat{P}_n$ is difficult (impossible) to estimate. The same then goes for the asymptotic distribution of $\Psi(\hat{P}_n)$.
- A targeted estimator, on the other hand, can converge at the parametric rate, and we can easily estimate the asymptotic distribution, even when it is constructed from estimators of the nuisance parameters that converge at a non-parametric rate.