EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

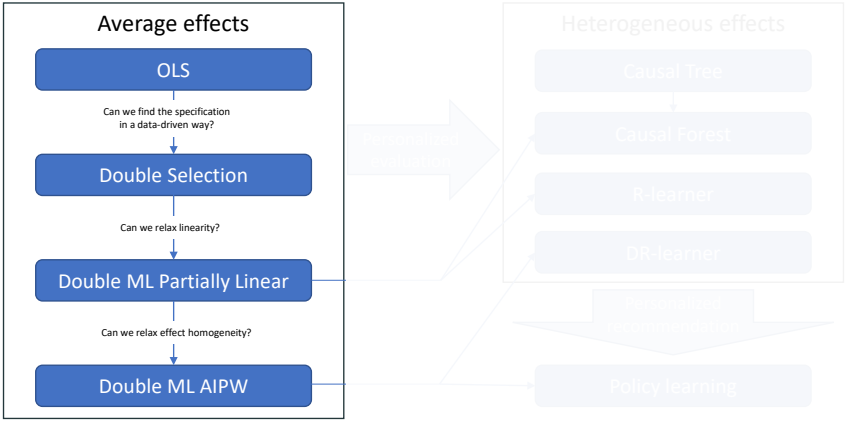
Causal Machine Learning

Double ML - the general recipe

Michael Knaus

WiSe 25/26

Current state of affairs



Double ML is a generic recipe and can be used for other target parameters/research designs

Plan of this morning

Understand the generic recipe of Double ML

1. The Double ML recipe
2. Average treatment effect on the treated
3. Instrumental variables
4. Standard errors with influence functions
5. More Double ML

The Double ML recipe

Some definitions

Let

- O be a collection of Observable variables, e.g. $O = (W, X, Y)$
- θ be the target parameter
- η be the collection of nuisance parameters, e.g. $\eta = (m(X), e(X))$

Double ML uses score functions $\psi(O; \tilde{\theta}, \tilde{\eta})$ that satisfy

1. $\overbrace{\mathbb{E}[\psi(O; \theta, \eta)] = 0}^{\text{moment condition}}$, i.e. with expectation zero if evaluated at true parameters
2. $\partial_r \, \mathbb{E}[\psi(O; \theta, \eta + r(\tilde{\eta} - \eta))]|_{r=0} = 0$, i.e. Neyman-orthogonality

Examples

Recall that the moment condition of the residual-on-residual regression with $m(X) := \mathbb{E}[Y \mid X]$ and $e(X) := \mathbb{E}[W \mid X]$ reads:

$$\mathbb{E}\left[(Y - m(X) - \tau(W - e(X)))(W - e(X))\right] = 0$$

$\Rightarrow O = (W, X, Y),\ \theta = \tau,\ \eta = (m(X), e(X))$

AIPW for ATE moment condition with $m(w, X) := \mathbb{E}[Y \mid W = w, X]$:

$$\mathbb{E}\left[m(1, X) - m(0, X) + \frac{W(Y - m(1, X))}{e(X)} - \frac{(1 - W)(Y - m(0, X))}{1 - e(X)} - \tau_{ATE}\right] = 0$$

$\Rightarrow O = (W, X, Y),\ \theta = \tau_{ATE},\ \eta = (m(1, X), m(0, X), e(X))$

Linear score functions

We will focus on linear score functions that can be represented as

$$\psi(O; \tilde{\theta}, \tilde{\eta}) = \tilde{\theta}\psi_a(O; \tilde{\eta}) + \psi_b(O; \tilde{\eta})$$

such that the moment condition can be written as

$$\mathbb{E}[\psi(O; \theta, \eta)] = \theta\, \mathbb{E}[\psi_a(O; \eta)] + \mathbb{E}[\psi_b(O; \eta)] = 0$$

and the solution is

$$\theta = -\frac{\mathbb{E}[\psi_b(O; \eta)]}{\mathbb{E}[\psi_a(O; \eta)]}$$

Example residual-on-residual regression

Moment condition:

$$\mathbb{E}\left[\big(Y - m(X) - \tau(W - e(X))\big)(W - e(X))\right] = 0$$

$$\mathbb{E}\left[(Y - m(X))(W - e(X)) - \tau(W - e(X))(W - e(X))\right] = 0$$

$$\tau\,\mathbb{E}[\underbrace{(-1)(W - e(X))^2}_{\psi_a}] + \mathbb{E}[\underbrace{(Y - m(X))(W - e(X))}_{\psi_b}] = 0$$

$$\Rightarrow \tau = -\frac{\mathbb{E}[\psi_b(O;\eta)]}{\mathbb{E}[\psi_a(O;\eta)]} = \frac{\mathbb{E}\left[(Y - m(X))(W - e(X))\right]}{\mathbb{E}\left[(W - e(X))^2\right]}$$

Example AIPW

AIPW for ATE moment condition:

$$\mathbb{E}\left[m(1,X) - m(0,X) + \frac{W(Y - m(1,X))}{e(X)} - \frac{(1-W)(Y - m(0,X))}{1-e(X)} - \tau_{ATE}\right] = 0$$

$$\tau_{ATE}\underbrace{(-1)}_{\psi_a} + \mathbb{E}\left[\underbrace{m(1,X) - m(0,X) + \frac{W(Y - m(1,X))}{e(X)} - \frac{(1-W)(Y - m(0,X))}{1-e(X)}}_{\psi_b}\right] = 0$$

$$\Rightarrow \tau_{ATE} = -\frac{\mathbb{E}[\psi_b(O;\eta)]}{\mathbb{E}[\psi_a(O;\eta)]} = \mathbb{E}\left[m(1,X) - m(0,X) + \frac{W(Y - m(1,X))}{e(X)} - \frac{(1-W)(Y - m(0,X))}{1-e(X)}\right]$$

This is a very complicated way to say that we take the expectation of the pseudo-outcome we called $\tilde{Y}_{ATE}$ last week, but it illustrates the recipe

Double ML recipe

1. Find Neyman-orthogonal score for your target parameter (can be constructed on demand, see Sec. 2 of Chernozhukov et al., 2018)
2. Predict nuisance parameters $\hat{\eta}$ with cross-fitted high-quality ML
3. Solve empirical moment condition to estimate the target parameter

$$\hat{\theta} = -\frac{\sum_i \psi_b(O_i; \hat{\eta}_i)}{\sum_i \psi_a(O_i; \hat{\eta}_i)}$$

4. Estimate standard error

$$\hat{\sigma}^2 = \frac{N^{-1} \sum_i \psi(O_i; \hat{\theta}, \hat{\eta}_i)^2}{[N^{-1} \sum_i \psi_a(O_i; \hat{\eta}_i)]^2} \Rightarrow \widehat{se}(\hat{\theta}) = \sqrt{\frac{\hat{\sigma}^2}{N}}$$

🤷 Don't panic, we will unpack this later, but first some use cases

Average treatment effect on the treated

Average treatment effect on the treated

Average Treatment Effect on the Treated (ATT): $\tau_{ATT} := \mathbb{E}[Y(1) - Y(0) \mid W = 1]$

- What is the expected treatment effect in the treated subpopulation?

Why is this an interesting parameter?

It helps us to evaluate the quality of treatment assignment if we compare it to the *ATE* (assuming higher outcomes are better):

- $ATT > ATE$: Treatment assignment better than random
- $ATT = ATE$: Treatment assignment as good as random
- $ATT < ATE$: Treatment assignment worse than random

See also very nice illustration of Andrew Heiss

ATT AIPW moment condition

This is the Neyman-orthogonal moment condition for *ATT*

$$\mathbb{E}\left[\overbrace{\underbrace{\frac{W}{e}(Y-m(0,X))}_{\text{regression adjustment}}-\underbrace{\frac{(1-W)e(X)}{e(1-e(X))}(Y-m(0,X))}_{\text{IPW weighted residual}}}^{\tilde{Y}_{ATT}}-\tau_{ATT}\frac{W}{e}\right]=0 \tag{1}$$

where $e := \mathbb{E}[W]$

$\Rightarrow O=(W,X,Y),\ \theta=\tau_{ATT},\ \eta=(m(0,X),e(X),e)$[1], $\psi_a=(-1)\frac{W}{e}, \psi_b=\tilde{Y}_{ATT}$

[1]The constant e is redundant (could multiply (1) by e to get rid of it), but it is there to fit into the generic variance estimation strategy below, see also Chernozhukov et al. (2018), Sec. 5.1

Instrumental variables

The framework: graph unconditional

A very popular research design (at least in economics) for identifying causal effects is to assume access to an instrumental variable

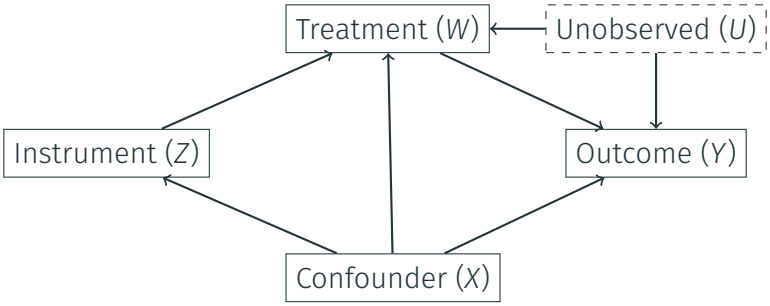
A valid instrument affects the outcome only through the treatment and we can use this exogenous treatment variation to identify its effect

Nice refresher by Kate Barnes

The framework: graph conditional

If Z is not 🎲, we still need to adjust for confounders

Let's check the simple case (our SWIG + d-separation skills would help with more complex scenarios)



Current workhorse for estimation is TSLS with same model selection issues as OLS

Partially linear IV moment condition

Robinson/partialling out style moment condition with Neyman-orthogonal score and new nuisance parameter $h(X) := \mathbb{E}[Z \mid X]$

$$\mathbb{E}\left[\Big(\overbrace{Y - m(X)}^{\text{outcome residual}} - \tau \overbrace{(W - e(X))}^{\text{treatment residual}} \Big) \overbrace{(Z - h(X))}^{\text{instrument residual}} \right] = 0$$

$$\mathbb{E}\left[(Y - m(X))(Z - h(X))) - \tau (W - e(X))(Z - h(X)) \right] = 0$$

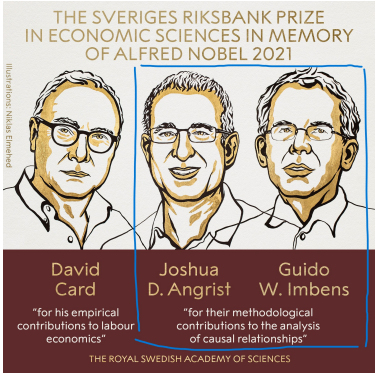
$$\tau \, \mathbb{E}[\underbrace{(-1)(W - e(X))(Z - h(X))}_{\psi_a}] + \mathbb{E}[\underbrace{(Y - m(X))(Z - h(X))}_{\psi_b}] = 0$$

$\Rightarrow O = (W, X, Y, Z),\ \theta = \tau,\ \eta = (m(X), e(X), h(X))$

The standard recipe applies (implemented in `DoubleML`)

Nobel price alert

What if effects are not homogeneous?



Angrist explanation

Imbens explanation

Local Average Treatment Effects (1/3)

In the case of randomized binary instrument Z and binary treatment W with potential treatments $W(Z)$, Imbens & Angrist (1994) show that the Wald estimator

$$\underbrace{\frac{\overbrace{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}^{\text{reduced form / intention to treat}}}{\underbrace{\mathbb{E}[W \mid Z=1] - \mathbb{E}[W \mid Z=0]}_{\text{first stage / complier share}}}}_{} = \mathbb{E}[Y(1) - Y(0) \mid \underbrace{W(1) - W(0) = 1}_{\text{complier}}] =: \tau_{LATE}$$

identifies a Local Average Treatment Effect (LATE) under monotonicity that nobody is moved out of the treatment by the instrument (no defiers)

The LATE describes the effect of the compliers who change their treatment status due to the instrument (may or may not be an interesting target parameter)

Local Average Treatment Effects (2/3)

What if Z is not randomly assigned, but we assume we observe a VAS for Z?

Double ML uses the following moment condition with a Neyman-orthogonal score

$$\mathbb{E}\Bigg[\overbrace{m_z(1,X) - m_z(0,X) + \frac{Z(Y - m_z(1,X))}{h(X)} - \frac{(1-Z)(Y - m_z(0,X))}{1-h(X)}}^{\psi_b}\Bigg] \tag{2}$$
$$+ \mathbb{E}\Bigg[\underbrace{(-1)\Big[e(1,X) - e(0,X) + \frac{Z(W - e(1,X))}{h(X)} - \frac{(1-Z)(W - e(0,X))}{1-h(X)}\Big]}_{\psi_a}\Bigg] \times \tau_{LATE} = 0$$

where $m_z(z,X) = \mathbb{E}[Y \mid Z = z, X]$, $h(X) = \mathbb{P}(Z = 1 \mid X)$ is the probability to be "instrumented" and $e(Z,X) = \mathbb{P}(W = 1 \mid Z, X)$ the probability to be treated given the instrument

Local Average Treatment Effects (3/3)

Equation (2) looks terrifying but leads to a familiar structure

$$\tau_{LATE} = \frac{\overbrace{\mathbb{E}\left[m_z(1,X) - m_z(0,X) + \frac{Z(Y - m_z(1,X))}{h(X)} - \frac{(1-Z)(Y - m_z(0,X))}{1-h(X)}\right]}^{\text{reduced form / intention to treat}}}{\underbrace{\mathbb{E}\left[e(1,X) - e(0,X) + \frac{Z(W - e(1,X))}{h(X)} - \frac{(1-Z)(W - e(0,X))}{1-h(X)}\right]}_{\text{first stage / complier share}}} \tag{3}$$

It generalizes the Wald estimator to the case with confounders

It just divides the ATE of the instrument on the outcome (reduced form) by the ATE of the instrument on the treatment (first stage)

See also Levis, Kennedy & Keele (2024) for excellent review of IV based identification and ML based estimation

Standard errors with influence functions

How to do statistical inference for Double ML

I told you that we estimate standard errors like this

$$\hat{\sigma}^2 = \frac{N^{-1} \sum_i \psi(O_i; \hat{\theta}, \hat{\eta}_i)^2}{[N^{-1} \sum_i \psi_a(O_i; \hat{\eta}_i)]^2} \Rightarrow \widehat{se}(\hat{\theta}) = \sqrt{\frac{\hat{\sigma}^2}{N}}$$

But why? 🤔

For better understanding, we need to introduce the concept of influence functions

Influence functions are powerful tools beyond Double ML, but we will focus on its use for our special case with linear scores and do not go into the technical details

For more general introductions see Kahn (2022) or Jann (2019, 2020)

Influence function

How is it defined?

$$\Psi(O;\theta,\eta) := -\,\mathbb{E}\left[\frac{\partial\psi}{\partial\theta}\right]^{-1}\psi(O;\theta,\eta) = -\,\mathbb{E}[\psi_a(O;\eta)]^{-1}\psi(O;\theta,\eta)$$

$\Rightarrow$ It is a scaled version of the score evaluated at the true parameter values

Important features:

- $\mathbb{E}[\Psi(O;\theta,\eta)] = \mathbb{E}[-\,\mathbb{E}[\psi_a(O;\eta)]^{-1}\psi(O;\theta,\eta)] = -\,\mathbb{E}[\psi_a(O;\eta)]^{-1}\underbrace{\mathbb{E}[\psi(O;\theta,\eta)]}_{=0} = 0$
- $\Psi(O_i;\theta,\eta_i)/N$ approximates the influence of observation i on the estimate of the target parameter

Influence function for inference (1/2)

What makes it so valuable for statistical inference?

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_i \Psi(O_i; \theta, \eta_i) + o_p(1) \xrightarrow{d} N(0, \underbrace{Var[\Psi(O; \theta, \eta)]}_{\sigma^2})$$

$\Rightarrow$ The estimator distribution and the influence function are closely linked #CLT

Note that (suppressing the arguments)

$$\sigma^2 = Var[\Psi] = \mathbb{E}[\Psi^2] - \underbrace{\mathbb{E}[\Psi]^2}_{=0} = \mathbb{E}[\Psi^2] = \mathbb{E}[(-\mathbb{E}[\psi_a]^{-1}\psi)^2] = \mathbb{E}[\psi_a]^{-2}\,\mathbb{E}[\psi^2] = \frac{\mathbb{E}[\psi^2]}{\mathbb{E}[\psi_a]^2}$$

Influence function for inference (2/2)

The sample equivalent is therefore

$$\hat{\sigma}^2 = \frac{N^{-1} \sum_i \psi(O_i; \hat{\theta}, \hat{\eta}_i)^2}{[N^{-1} \sum_i \psi_a(O_i; \hat{\eta}_i)]^2}$$

and the standard error is estimated as $\widehat{se}(\hat{\theta}) = \sqrt{\frac{\hat{\sigma}^2}{N}}$

$\widehat{se}(\hat{\theta})$ can then be used to calculate *t*-values, *p*-values, confidence intervals etc.

Example *ATE*

Start with the ATE score where we denote $m(W, X) = m_W$ and $e(X) = e$

$$\psi_{ATE} = \overbrace{\underbrace{m_1 - m_0 + \frac{(D-e)(Y-m_W)}{e(1-e)}}_{\psi_b}}^{\tilde{Y}_{ATE}} + \underbrace{(-1)}_{\psi_a} \tau_{ATE}$$

$$\Rightarrow \frac{\partial \psi_{ATE}}{\partial \tau_{ATE}} = \psi_a = -1$$

$$\Rightarrow \Psi_{ATE} = -\mathbb{E}[(-1)]^{-1}[\tilde{Y}_{ATE} - \tau_{ATE}] = \tilde{Y}_{ATE} - \tau_{ATE}$$

In the special case where $\mathbb{E}[\psi_a] = -1$, score and influence function coincide

Influence function for inference: Chain rule

A cool/convenient feature of influence functions is that they obey the chain rule

Imagine that the target parameter is a function of k other parameters, i.e. $\theta = f(\theta_1, ..., \theta_K)$

Then the influence function of θ is

$$\Psi_\theta = \sum_{k=1}^{K} \frac{\partial f}{\partial \theta_k} \Psi_{\theta_k}$$

$\Rightarrow$ we can use existing influence functions to create new ones

$\Rightarrow$ very powerful way to get standard errors of complicated objects

Example $ATT - ATE$

We may want to test whether the $ATT = ATE$

For this purpose, we create the new parameter $\Delta(\tau_{ATT}, \tau_{ATE}) = \tau_{ATT} - \tau_{ATE}$

The new influence function is

$$\begin{aligned}
\Psi_\Delta &= \overbrace{\frac{\partial \Delta}{\partial \tau_{ATT}}}^{=1} \Psi_{\tau_{ATT}} + \overbrace{\frac{\partial \Delta}{\partial \tau_{ATE}}}^{=-1} \Psi_{\tau_{ATE}} \\
&= \Psi_{\tau_{ATT}} - \Psi_{\tau_{ATE}} \\
&= \tilde{Y}_{ATT} - \tau_{ATT} W/e - \tilde{Y}_{ATE} + \tau_{ATE}
\end{aligned}$$

and can be used to get $\widehat{se}(\hat{\Delta})$

Two ways to get the *LATE* influence function: (1) direct way

The direct way starts with the score implied by (2) where we denote $m(Z,X) = m_Z$, $e(Z,X) = e_Z$, $h(X) = h$, and rewrite the weighted residuals for compactness

$$\psi_{LATE} = \underbrace{\overbrace{m_1 - m_0 + \frac{(Z-h)(Y-m_Z)}{h(1-h)}}^{\tilde{Y}_{Z\to Y}}}_{\psi_b} - \underbrace{\left[\overbrace{e_1 - e_0 + \frac{(Z-h)(Y-e_Z)}{h(1-h)}}^{\tilde{Y}_{Z\to W}}\right]}_{\psi_a} \times \tau_{LATE}$$

$$\Rightarrow \frac{\partial \psi_{LATE}}{\partial \tau_{LATE}} = \psi_a = -\tilde{Y}_{Z\to W}$$

$$\begin{aligned}\Rightarrow \Psi_{LATE} &= -\mathbb{E}[-\tilde{Y}_{Z\to W}]^{-1}[\tilde{Y}_{Z\to Y} - \tilde{Y}_{Z\to W} \times \tau_{LATE}] \\ &= \mathbb{E}[\tilde{Y}_{Z\to W}]^{-1}[\tilde{Y}_{Z\to Y} - \tilde{Y}_{Z\to W} \times \tau_{LATE}]\end{aligned}$$

Two ways to get the *LATE* influence function: (2) chain rule

Denote by $\tau_{Z\to W}$ and $\tau_{Z\to Y}$ the ATEs of Z on W and Y, respectively

We start by noting that

$$\tau_{LATE}(\tau_{Z\to Y}, \tau_{Z\to W}) = \frac{\tau_{Z\to Y}}{\tau_{Z\to W}}$$

The chain rule tells us that

$$\begin{aligned}
\Psi_{LATE} &= \frac{\partial \tau_{LATE}}{\partial \tau_{Z\to Y}}\Psi_{\tau_{Z\to Y}} + \frac{\partial \tau_{LATE}}{\partial \tau_{Z\to W}}\Psi_{\tau_{Z\to W}} = \frac{1}{\tau_{Z\to W}}\Psi_{\tau_{Z\to Y}} - \frac{\tau_{Z\to Y}}{\tau^2_{Z\to W}}\Psi_{\tau_{Z\to W}} \\
&= \frac{1}{\tau_{Z\to W}}(\tilde{Y}_{Z\to Y} - \tau_{Z\to Y}) - \frac{\tau_{Z\to Y}}{\tau^2_{Z\to W}}(\tilde{Y}_{Z\to W} - \tau_{Z\to W}) \\
&= \frac{1}{\tau_{Z\to W}}\left[\tilde{Y}_{Z\to Y} - \cancel{\tau_{Z\to Y}} - \underbrace{\frac{\tau_{Z\to Y}}{\tau_{Z\to W}}}_{\tau_{LATE}}\tilde{Y}_{Z\to W} + \cancel{\frac{\tau_{Z\to Y}}{\tau_{Z\to W}}\tau_{Z\to W}}\right] \\
&= \mathbb{E}[\tilde{Y}_{Z\to W}]^{-1}[\tilde{Y}_{Z\to Y} - \tilde{Y}_{Z\to W} \times \tau_{LATE}]
\end{aligned}$$

b/c $\tau_{Z\to W} = \mathbb{E}[\tilde{Y}_{Z\to W}]$

Simulation notebook: Influence functions explained using OLS

Application notebook: Double ML as generic recipe

More Double ML

The generic concept is applied to many other scenarios

Difference-in-differences: Chang (2020), Zimmert (2020), and Sant'Anna and Zhao (2020), Nie, Lu and Wager (2019) or Hatamyar et al. (2023) (see also `DoubleML` for an implementation)

Regression discontinuity: Noack, Olma and Rothe (2024)

Mediation: Farbmacher et al. (2022)

Dynamic treatments: Bodory et al. (2022)

Quantile treatment effects: Belloni et al. (2017) and Kallus, Mao & Uehara (2019), Baklicharov et al. (2024)

and many more (to come)

Important for you is that they all build on principles you know now

More about the general recipe

Ahrens, Chernozhukov, Hansen, Kozbur, Schaffer & Wiemann (2025) provide a nice introduction

Ahrens, Hansen, Schaffer & Wiemann (2025) emphasize the role of using stacked/ensemble/super-learner in the Double ML framework

Ceterum censeo a fancy method alone is not a credible identification strategy

$\Rightarrow$ separate identification and estimation