EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Causal Machine Learning

Estimating constant effects: Double Selection to Double ML

Michael Knaus

WiSe 23/24

Plan of this morning

How can we leverage supervised ML to adjust for confounding?

1. Helpful and familiar concepts

2. The identification strategy with many names

3. Causal inference in linear models

4. Double Selection

5. Double ML: Partially linear model

Helpful and familiar concepts

Focus on what we want to know (1/2)

Causal ML methods force us to distinguish between two types of parameters:

1. **Target parameter** is motivated by the research question and defined under modelling assumption, usually it is only one- or at least low-dimensional (e.g. effect of policy on sth.)
2. **Nuisance parameters** are inputs that are required to get our hands on the target parameter, but are not relevant for our research question

I think this is really healthy as we commit to the target parameter and are not tempted to interpret every single coefficient in a regression output #Table2Fallacy

Focus on what we want to know (2/2)

You know this concept from instrumental variables

We use instrumental variables if we are concerned about endogeneity with regard to a particular variable of interest (omitted variables or reverse causality)

The parameter of this variable is our target parameter and the fitted values of the first stage in 2SLS is the nuisance parameter

Usually nobody is interested in the fitted values of the first stage

However, we need them to consistently estimate the target parameter in the second stage

Rewriting stuff

You will see that Causal ML is mostly about rewriting stuff such that we are allowed to leverage supervised ML to estimate the nuisance parameters

Importantly, the parameters of interest remain the same although the rewritten form can look quite different to the original/familiar model

The methods usually boil down to running multiple supervised ML regressions and combining their predictions into a final OLS regression

The crucial point is that the statistical inference in this final stage is valid if we follow a particular recipe

$\Rightarrow$ You will learn how to split the estimation of causal effects into prediction tasks

Frisch-Waugh-Theorem

This may sound familiar from the Frisch-Waugh-Theorem (FWT)

The FWT tells us that we can estimate θ in a standard linear regression $Y = W\theta + X'\beta + U$ in a three-stage procedure:

1. Run a regression of the form $Y = \alpha_y + X'\pi + U_{Y\sim X}$ and extract the estimated residuals $\hat{U}_{Y\sim X}$
2. Run a regression of the form $W = \alpha_w + X'\delta + U_{W\sim X}$ and extract the estimated residuals $\hat{U}_{W\sim X}$
3. Run a residual-on-residual regression of the form $\hat{U}_{Y\sim X} = \theta\hat{U}_{W\sim X} + \epsilon$

The resulting estimate $\hat{\theta}$ is numerically identical to the estimate we would get if we just run the full OLS model

Simulation Notebook: Basics: OLS and Frisch-Waugh

Panel methods

Also in panel settings it is common to transform the problem without altering the target parameter

Fixed-effects regression

Consider a linear panel model with unobserved individual fixed-effect:

$$Y_{it} = W_{it}\theta + X_{it}\beta + \alpha_i + \epsilon_{it}$$

Demeaning transformation:

$$\underbrace{Y_{it} - \bar{Y}_i}_{\text{pseudo-outcome}} = \underbrace{(W_{it} - \bar{W}_i)}_{\text{pseudo-treatment}}\ \theta + \underbrace{(X_{it} - \bar{X}_i)}_{\text{pseudo-covariates}}\ \beta + (\epsilon_{it} - \bar{\epsilon}_i)$$

$\Rightarrow$ The problem is transformed and boils down to running OLS with pseudo-outcomes/-treatments/-covariates, but the θ is unaffected

The identification strategy with many names

Adjusting for confounding factors

Especially in Social Science, randomization of treatments is often not possible

$\Rightarrow$ We need ways to work with so-called observational data

The strategy where causal ML has the most obvious benefit aims to adjust for all confounding factors and is known under many different names:

- Backdoor adjustment
- Conditional independence assumption
- Exchangeability
- Exogeneity
- Ignorability
- Measured confounding
- No unmeasured confounding
- Selection-on-observables
- Unconfoundedness
- ...

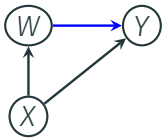
Measured confounding - SCM and DAG

For now I will go with measured confounding because it is a direct and short description of what is required

We assume the structural causal model

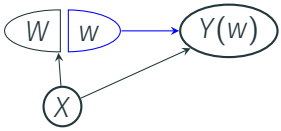
$$
\begin{aligned}
X &= f_X(U_X) \\
W &= f_W(X, U_W) \\
Y &= f_Y(W, X, U_Y)
\end{aligned}
$$

represented as the following DAG



Measured confounding - SWIG

The SWIG that produces potential outcomes looks like this



and implies

Identifying Assumption 1 (measured confounding)

$$Y(w) \perp\!\!\!\perp W \mid X \text{ for all } w \in \mathcal{W} \subset \mathbb{R} \tag{1}$$

where we allow in this slide deck that the treatment is continuous

See identification notebook "DAG and SWIG for measured confounding" for an R assisted derivation

Measured confounding - confounder selection

Credible identification demands to define the set of variables that make assumption (1) credible to hold

This requires good knowledge about the treatment assignment mechanism and can be guided by the machinery developed around DAGs

In particular, they provide a principled framework to disentangle good and bad controls (see, e.g. Cinelly, Forley and Pearl, 2022, for a crash course)

However, our focus is on estimation

This means we start after we obtained a set of variables X that allow credible identification and innovate in the estimation part

Causal inference in linear models

Starting with linear model

Probably the first contact you had with such settings is within linear models that are estimated via OLS $\Rightarrow$ we also start there

To identify effects in a linear model, we impose a functional form assumption

Modelling Assumption 1 (Linear potential outcomes)

Potential outcomes are linear functions of confounding variables X:

$$Y(w) = \tau w + X'\beta + U_{Y(w)}; \;\; \mathbb{E}[U_{Y(w)} \mid X] = 0; \;\forall\, w \in \mathcal{W}$$

$\Rightarrow$ The CEF of the potential outcome is $\mathbb{E}[Y(w) \mid X] = \tau w + X'\beta$

Parameters

This model implies relatively boring effects of a one-unit increase in W:

$$\begin{aligned} ITE &= Y(w+1) - Y(w) = \tau(w+1) + X'\beta + U_{Y(w+1)} - \tau w - X'\beta - U_{Y(w)} \\ &= \tau + U_{Y(w+1)} - U_{Y(w)} \end{aligned}$$

$$\begin{aligned} CATE(X) &= \mathbb{E}[\tau + U_{Y(w+1)} - U_{Y(w)} \mid X] = \tau + \mathbb{E}[U_{Y(w+1)} \mid X] - \mathbb{E}[U_{Y(w)} \mid X] \overset{MA1}{=} \tau \\ ATE &= \mathbb{E}[\mathbb{E}[\tau \mid X]] = \tau \end{aligned}$$

$\Rightarrow$ MA1 implies constant CATEs $\Rightarrow$ constant average effects

$\Rightarrow$ We assume effect homogeneity

$\Rightarrow \tau$ is our target parameter (definition ✅), but not yet identified

Identification under linear model (1/2)

IA1 implies that $\mathbb{E}[Y(w) \mid X] = \mathbb{E}[Y(w) \mid W, X]$, i.e. that the potential outcome is conditionally mean independent of the treatment

Plugging in MA1 we observe that

$$\tau w + X'\beta + \mathbb{E}[U_{Y(w)} \mid X] = \tau w + X'\beta + \mathbb{E}[U_{Y(w)} \mid W, X]$$

$\Rightarrow$ IA1 in combination with MA1 implies

$$\mathbb{E}[U_{Y(w)} \mid X] = \mathbb{E}[U_{Y(w)} \mid W, X] = 0 \tag{2}$$

Identification under linear model (2/2)

Under consistency, the observed outcome is

$$Y = Y(W) = \tau W + X'\beta + U_{Y(W)}$$

The CEF of the observed outcome is therefore

$$\mathbb{E}[Y \mid W, X] = \tau W + X'\beta + \mathbb{E}[U_{Y(W)} \mid W, X] \overset{(2)}{=} \tau W + X'\beta$$

$\Rightarrow$ The unknown parameters τ, β can be obtained as solution of a least squares problem using only observable variables:

$$(\tau, \beta) = \underset{\check{\tau}, \check{\beta}}{\arg\min}\, \mathbb{E}[(Y - \check{\tau}W - X'\check{\beta})^2] \tag{3}$$

$\Rightarrow$ The effect τ is a function of observable variables $\Rightarrow$ identified ✅

Bonus: Potential confusions

I always got confused by unobservables, error terms and residuals in linear models and OLS because they often receive the same letter to discuss identification and estimation

Note that writing an outcome model like $Y = \theta W + X'\beta + \varepsilon;\ \mathbb{E}[\varepsilon \mid W, X] = 0$ only implies that the CEF is linear and does not guarantee that θ coincides with the causal parameter τ

It does not rule out, e.g., that $\mathbb{E}[U_{Y(W)} \mid W, X] = \beta_u W$ such that CEF of observed outcome could indeed be linear but would "identify" a non-causal parameter even under MA1:

$$\begin{aligned}\mathbb{E}[Y(W) \mid W, X] &\overset{Cons.}{=} \mathbb{E}[Y \mid W, X] \overset{MA1}{=} \tau W + X'\beta + \mathbb{E}[U_{Y(W)} \mid W, X] \\ &= \tau W + X'\beta + \beta_u W = \underbrace{(\tau + \beta_u)}_{=\theta} W + X'\beta\end{aligned}$$

$\beta_u \neq 0$ means that W correlates with unobserved factors $\Rightarrow$ omitted variable bias

This is not ruled out by simply writing $\mathbb{E}[\varepsilon \mid W, X] = 0$ on the observed outcome level $\Rightarrow$ We need to assume conditional independence on the potential outcomes level

See this little note for further discussion

Double Selection

From identification to estimation via OLS

The identification result motivates the estimation via OLS as the sample analog of population problem (3):

$$\left(\hat{\tau}, \hat{\beta}\right) = \underset{\check{\tau},\check{\beta}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \check{\tau} W_i - X_i' \check{\beta}\right)^2 \tag{4}$$

$\Rightarrow \hat{\tau}$ is our estimated homogeneous treatment effect...

... but wait ...

we have committed ourselves to a set of variables that make IA1 plausible to hold, but how *exactly* should they enter the estimation process? 🤷

Challenge: model selection

A classic: Age

Often we want to adjust for age. But then it begins, *"should it enter linearly, with a quadratic term, or even a cubic, ..., or WAIT, maybe we should discretize age and include it as categorical variables, hmm, but should we use then age groups of 4 or 5 years, ..., and should we interact them with the female dummy, ..."* 🤯

These are painful decisions

The final model is then often the result of a more or less principled process based on t-tests, F-tests, (adjusted) R^2 or other magic tools

Remark: We do not even need high-dimensional controls to run into these issues, one continuous variable suffices

The journey begins

Double Selection: Procedure

Belloni et al. (2014) propose the **Double Selection** procedure:

1. Assume approximate sparsity (discussed below)
2. Select variables that predict outcome via Post-Lasso (w/o treatment variable)
3. Select variables that predict treatment via Post-Lasso
4. Use union of selected variables w/ treatment variable in OLS

This procedure constructively addresses the impossibility results of valid inference after model selection if we are committed to a target parameter

Double Selection: motivation

Why is Double Selection necessary and single selection problematic?

We can use our well trained understanding of Omitted Variable Bias (OVB)

OVB occurs if

(i) the OV has a non-zero coefficient in the outcome model
AND
(ii) the OV is correlated with the treatment variable W

Variable selection based on outcome model does not incorporate (ii) because it cares about prediction MSE and is not aware that we care about unbiasedness of one particular parameter

$\Rightarrow$ Variables with relatively small coefficient in the outcome model might be ignored although they are highly predictive for the treatment $\Rightarrow$ OVB

A taxonomy of (potential) control variables

Consider the two regression equations underlying Double Selection:

$$Y = \alpha_y + \pi X + U_{Y \sim X} \tag{5}$$

$$W = \alpha_w + \delta X + U_{W \sim X} \tag{6}$$

To make the point without unnecessary notation assume that X is standardized

Then, we can categorize the variables in X according to their coefficients:

		π		
		Zero	**Small**	**Large**
δ	**Zero**	Irrelevant	Irrelevant	Irrelevant
	Small	Irrelevant	Greyzone	Important
	Large	Irrelevant	Important	Essential

Single selection with outcome equation could miss the small π, large δ group

Derivation of Y model

$$
\begin{aligned}
Y &= \tau W + \beta X + U_{Y\sim W+X} = \tau(\alpha_W + \delta X + U_{W\sim X}) + \beta X + U_{Y\sim W+X} \\
&= \underbrace{\tau\alpha_W}_{\alpha_y} + \underbrace{(\beta + \tau\delta)}_{=\pi} X + \underbrace{\tau U_{W\sim X} + U_{Y\sim W+X}}_{=U_{Y\sim X}} = \alpha_y + \pi X + U_{Y\sim X}
\end{aligned}
$$

Approximate sparsity

Approximate sparsity

The number of controls with non-zero coefficients in (5) and (6) is small relative to the sample size (for technical definition, see Sec 2.1 of Belloni et al., 2014).

Resembles practice to specify relatively sparse models

But allows to remain agnostic about the model and to "let the data speak"

Most importantly standard robust OLS standard errors are valid even after model selection

For me this is such a large benefit at basically no costs (fast implementation `hdm`)

Discussion Double Selection

Advantages:

- Easy to implement, statistically valid and transparent way of model selection
- Easy to explain to an OLS focused audience
- Works for binary and continuous treatments

Disadvantages:

- We still have to commit to a set of potential controls

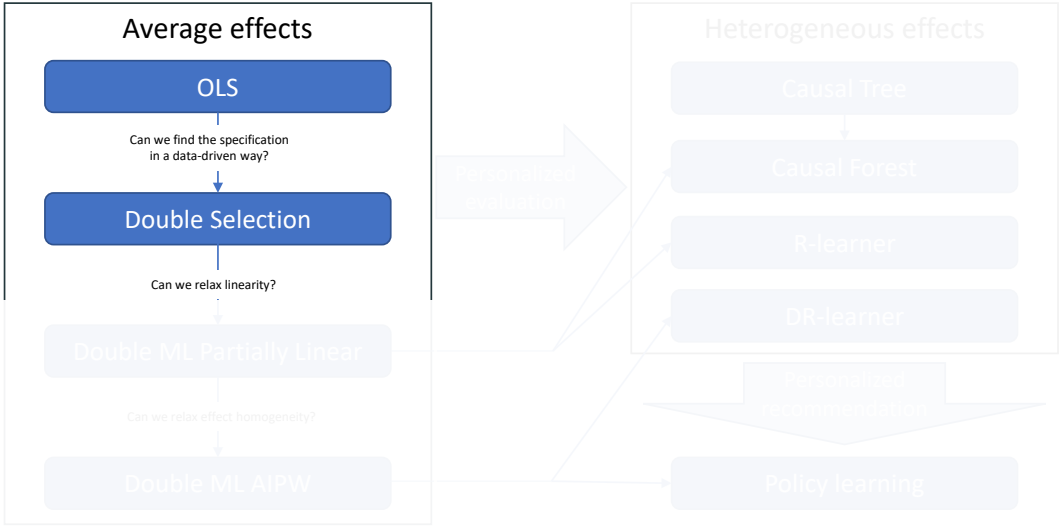
  $\Rightarrow$ which transformations of age reloaded
- Effect homogeneity often implausible
- Requires a linear potential outcome model and rules out non-linearities

Simulation notebook: Double selection

Application notebook: Double Selection and Partially Linear Double ML (part 1)

Double ML: Partially linear model

The journey continues



Relaxing the modelling assumption

The partially linear model relaxes the linearity assumption

Modelling Assumption 2 (Partially linear potential outcomes)

Potential outcomes are a partially linear function of confounding variables X:

$$Y(w) = \tau w + g(X) + U_{Y(w)};\ \ \mathbb{E}[U_{Y(w)} \mid X] = 0;\ \forall\, w \in \mathcal{W}$$

$\Rightarrow$ The CEF of the potential outcome is $\mathbb{E}[Y(w) \mid X] = \tau w + g(X)$

Parameters

Same as in the linear model but allowing for more flexible underlying POs

$$\begin{aligned} ITE &= Y(w+1) - Y(w) = \tau(w+1) + g(X) + U_{Y(w+1)} - \tau w - g(X) - U_{Y(w)} \\ &= \tau + U_{Y(w+1)} - U_{Y(w)} \end{aligned}$$

$$\begin{aligned} CATE(X) &= \mathbb{E}[\tau + U_{Y(w+1)} - U_{Y(w)} \mid X] = \tau + \mathbb{E}[U_{Y(w+1)} \mid X] - \mathbb{E}[U_{Y(w)} \mid X] \overset{MA2}{=} \tau \\ ATE &= \mathbb{E}[\mathbb{E}[\tau \mid X]] = \tau \end{aligned}$$

$\Rightarrow$ MA2 implies constant CATEs $\Rightarrow$ constant average effects

$\Rightarrow$ We assume effect homogeneity

$\Rightarrow \tau$ is our target parameter (definition ✅)

Identification under partial linearity (1/2)

Similar to the linear model the combination of IA1 and MA2 ensures that

$$\mathbb{E}[Y \mid W, X] = \tau W + g(X) + \mathbb{E}[U_{Y(W)} \mid W, X] = \tau W + g(X)$$

and the observed outcome is

$$Y = \tau W + g(X) + U_{Y(W)} \tag{7}$$

Identification under partial linearity (2/2)

Following Robinson (1988), we can rewrite (7) as

$$\underbrace{Y - \overbrace{\mathbb{E}[Y \mid X]}^{m(X)}}_{\text{outcome residual}} = \tau \underbrace{(W - \overbrace{\mathbb{E}[W \mid X]}^{e(X)})}_{\text{treatment residual}} + U_{Y(W)} \tag{8}$$

$\Rightarrow \tau$ is identified ✅ by a residual-on-residual regression w/o constant:

$$\tau = \arg\min_{\check{\tau}} \mathbb{E}[(\underbrace{Y - m(X)}_{\text{pseudo-outcome}} - \check{\tau} \underbrace{(W - e(X))}_{\text{single regressor}})^2] = \frac{Cov[W - e(X), Y - m(X)]}{Var[W - e(X)]} \tag{9}$$

You can think of this as a generalization of Frisch-Waugh

Derivation of residualized representation

Note that the CEF of the outcome given X is

$$\begin{aligned} m(X) := \mathbb{E}[Y \mid X] &= \mathbb{E}[\tau W + g(X) + U_{Y(W)} \mid X] \\ &= \tau e(X) + g(X) + \underbrace{\mathbb{E}[U_{Y(W)} \mid X]}_{\substack{MA2 \\ =0}} \\ &= \tau e(X) + g(X) \end{aligned} \tag{10}$$

Then, rewrite (7) as

$$\begin{aligned} Y &= \tau W + g(X) + U_{Y(W)} && \mid -m(X) \\ Y - m(X) &= \tau W + g(X) + U_{Y(W)} - (\tau e(X) + g(X)) \\ Y - m(X) &= \tau(W - e(X)) + U_{Y(W)} \end{aligned}$$

Identification in partially linear model

Suppressing dependencies of functions on X and using results from the previous slide, we can show

$$\begin{aligned}
\frac{Cov[W-e, Y-m]}{Var[W-e]} &\overset{(7),(10)}{=} \frac{Cov[W-e, \tau W + g + U_{Y(W)} - \tau e - g]}{Var[W-e]} \\
&= \frac{Cov[W-e, \tau(W-e) + U_{Y(W)}]}{Var[W-e]} \\
&= \tau \frac{Cov[W-e, W-e]}{Var[W-e]} + \frac{Cov[W-e, U_{Y(W)}]}{Var[W-e]} \\
&= \tau + \frac{Cov[W, U_{Y(W)}]}{Var[W-e]} - \frac{Cov[e, U_{Y(W)}]}{Var[W-e]} \\
&\overset{IA1}{=} \tau
\end{aligned}$$

b/c $Cov[W, U_{Y(W)}] = \mathbb{E}[W U_{Y(W)}] - \mathbb{E}[W] \underbrace{\mathbb{E}[U_{Y(W)}]}_{=0} \overset{LIE}{=} \mathbb{E}[\mathbb{E}[W U_{Y(W)} \mid W, X]] = \mathbb{E}[W \underbrace{\mathbb{E}[U_{Y(W)} \mid W, X]}_{\overset{IA1}{=}0}] = 0$, same for

$Cov[e, U_{Y(W)}] = 0$

From identification to estimation via Double ML

The identification result (9) would motivate the following estimator:

$$\hat{\tau}^{oracle} = \arg\min_{\check{\tau}} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - m(X_i) - \check{\tau}(W_i - e(X_i)) \right)^2 \tag{11}$$

where $m(X) := \mathbb{E}[Y \mid X]$ and $e(X) := \mathbb{E}[W \mid X]$ are outcome and treatment CEFs, the so-called nuisance parameters

BUT we usually do not know the outcome CEF and the treatment CEF

$\Rightarrow$ (11) is not feasible

$\Rightarrow$ We need to approximate $m(X)$ and $e(X)$

$\Rightarrow$ The ML toolbox might be helpful

Double ML for partially linear model: procedure

Chernozhukov et al. (2018) propose a three step procedure:

1. Form prediction model for the treatment: $\hat{e}(X)$
2. Form prediction model for the outcome: $\hat{m}(X)$
3. Run feasible residual-on-residual regression:

$$\hat{\tau} = \underset{\check{\tau}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \check{\tau}(W_i - \hat{e}(X_i)) \right)^2 \tag{12}$$

$\Rightarrow \hat{\tau}$ is our estimated homogeneous treatment effect $\Rightarrow$ Estimation ✅

But there is a catch, right? 🤨

Double ML for partially linear model: conditions

Nothing too serious, but the predictions of nuisance parameters $\hat{e}(X)$ and $\hat{m}(X)$ have to fulfill two conditions:

1. **High-quality**: Predictors must be consistent and have convergence rates of faster than $N^{1/4}$
2. **Out-of-sample**: Predictions of individual observation formed without the observation itself

Standard (robust) OLS inference for (12) is valid under these conditions (Chernozhukov et al. (2018), Theorem 4.1)

But what do these conditions mean? 🤔

High-quality predictions (1/2)

Consistency:

The ML methods converge to the true CEFs as $N \to \infty$

Convergence rate: (check R Notebook "Convergence rates" as a refresher)

Parametric models like OLS converge at the rate $N^{1/2}$ if the model of the CEF is correct

Their RMSE $\left(\mathbb{E}\left[\sqrt{(\hat{m}(X) - m(X))^2}\right]\right)$ is expected to halve if we increase sample size by factor four

ML methods usually do not converge as quickly because they can not leverage the structural information of a parametric model

For Double ML to work, it suffices that the RMSE more than halves if we increase sample size by factor 16 ($N^{1/4}$ convergence)

High-quality predictions (2/2)

Several popular ML methods can achieve the required convergence rates, e.g.:

- (Post-)Lasso (Belloni & Chernozukov, 2013)
- Random Forests (Wager & Walther, 2015; Syrgkanis & Zampetakis, 2020)
- Neural Networks (Farrell et al., 2021)

Each method requires different assumptions (mostly some form of sparsity), which (at least so far) can not be tested

$\Rightarrow$ Some structure needed, but less structure than imposed by parametric models

Cross-fitting

The second condition is easily satisfied by using K-fold cross-fitting

Example 2-fold cross-fitting:

- Randomly split the sample in two parts S^1 and S^2
- Learn prediction models $\hat{m}^1(x)$ and $\hat{e}^1(x)$ in S^1 and predict in S^2
- Learn prediction models $\hat{m}^2(x)$ and $\hat{e}^2(x)$ in S^2 and predict in S^1
- Use the cross-fitted predictions in residual-on-residual regression (12)

This ensures that the nuisance parameters induce no bias by overfitting (Chernozhukov et al. (2018), Section 1.1)

and that we waste no information

Why does it work?

The residual-on-residual regression has a so-called Neyman-orthogonal score

The score defines the solution of the minimization problem (12) (derivation wrt τ):

$$\frac{1}{N}\sum_{i=1}^{N}\underbrace{\big(Y_i - \hat{m}(X_i) - \hat{\tau}(W_i - \hat{e}(X_i))\big)(W_i - \hat{e}(X_i))}_{\psi(Y_i, W_i, \hat{\tau}, \hat{m}(X_i), \hat{e}(X_i))} = 0 \tag{13}$$

$$\Rightarrow \quad \hat{\tau} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{m}(X_i))(W_i - \hat{e}(X_i))}{\frac{1}{N}\sum_{i=1}^{N}(W_i - \hat{e}(X_i))^2} \tag{14}$$

Neyman-orthogonality of score $\psi(Y, W, \tau, m(X), e(X))$ ensures that the estimated $\hat{\tau}$ is immunized against small errors in the estimation of nuisance parameters

Remark: Using (7) directly would not be Neyman-orthogonal

Neyman-orthogonality of residual-on-residual regression

Neyman-orthogonality means that the Gateaux derivative with respect to the nuisance parameters is zero in expectation at the true nuisance parameters (NP)

$$\partial_r \mathbb{E}[\psi(..., m + r(\tilde{m} - m), e + r(\tilde{e} - e))]|_{r=0} = 0 \tag{15}$$

where we suppress the dependence of NP on X and denote by, e.g., $\tilde{m}$ a value of the outcome nuisance that is different from the true value m

This looks very scary, but we only need to know how to setup the problem and then take standard derivatives (you passed Microeconomics, right)

For simplictly, we get rid of the brackets and write the score with true target and nuisance parameters as

$$\psi(Y, W, \tau, m(X), e(X)) = \big(Y - m - \theta(W - e)\big)(W - e)$$

First, add perturbations to the true nuisance parameters in the score

$$\begin{aligned}
&\psi(..., m + r(\tilde{m} - m), e + r(\tilde{e} - e)) \\
&= (Y - [m + r(\tilde{m} - m)] - \theta(W - [e + r(\tilde{e} - e)])(W - [e + r(\tilde{e} - e)]) \\
&= (Y - [m + r(\tilde{m} - m)])(W - [e + r(\tilde{e} - e)]) - \theta(W - [e + r(\tilde{e} - e)])^2
\end{aligned}$$

Second, take the derivative wrt r (we can interchange differentiation and expectation due to the Leibniz integral rule)

$$\begin{aligned}
&\partial_r \psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e)) \\
&\quad = \partial_r \big[(Y - [m + r(\tilde{m} - m)])(W - [e + r(\tilde{e} - e)]) - \theta(W - [e + r(\tilde{e} - e)])^2\big] \\
&\quad = \partial_r \big[YW - Y[e + r(\tilde{e} - e)] - [m + r(\tilde{m} - m)]W + [m + r(\tilde{m} - m)][e + r(\tilde{e} - e)] \\
&\qquad\quad - \theta(W - [e + r(\tilde{e} - e)])^2\big] \\
&\quad = \partial_r \big[YW - Y[e + r(\tilde{e} - e)] - [m + r(\tilde{m} - m)]W + me + mr(\tilde{e} - e) + r(\tilde{m} - m)e \\
&\qquad\quad + r^2(\tilde{m} - m)(\tilde{e} - e) - \theta(W - [e + r(\tilde{e} - e)])^2\big] \\
&\quad = -Y(\tilde{e} - e) - (\tilde{m} - m)W + m(\tilde{e} - e) + (\tilde{m} - m)e + 2r(\tilde{m} - m)(\tilde{e} - e) \\
&\qquad\quad - 2\theta(W - [e + r(\tilde{e} - e)])(\tilde{e} - e)
\end{aligned}$$

Third, evaluate at $r = 0$

$$\begin{aligned}
&\partial_r \psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e))|_{r=0} \\
&\quad = -Y(\tilde{e} - e) - (\tilde{m} - m)W + m(\tilde{e} - e) + (\tilde{m} - m)e + 2 \cdot 0(\tilde{m} - m)(\tilde{e} - e) \\
&\qquad - 2\theta(W - [e + 0(\tilde{e} - e)])(\tilde{e} - e) \\
&\quad = -Y(\tilde{e} - e) - (\tilde{m} - m)W + m(\tilde{e} - e) + (\tilde{m} - m)e - 2\theta(W - e)(\tilde{e} - e)
\end{aligned}$$

Fourth, take expectation

$$\begin{aligned}
&\partial_r \, \mathbb{E}[\psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e))]|_{r=0} \\
&\quad = \mathbb{E}\left[-Y(\tilde{e} - e) - (\tilde{m} - m)W + m(\tilde{e} - e) + (\tilde{m} - m)e - 2\theta(W - e)(\tilde{e} - e)\right] \\
&\quad \overset{LIE}{=} \mathbb{E}\left[\mathbb{E}\left[-Y(\tilde{e} - e) - (\tilde{m} - m)W + m(\tilde{e} - e) + (\tilde{m} - m)e - 2\theta(W - e)(\tilde{e} - e)|X\right]\right] \\
&\quad = \mathbb{E}[\cancel{-m(\tilde{e} - e)} - \cancel{(\tilde{m} - m)e} + \cancel{m(\tilde{e} - e)} + \cancel{(\tilde{m} - m)e} - 2\theta \underbrace{(e - e)}_{=0}(\tilde{e} - e)] = 0
\end{aligned}$$

$\Rightarrow$ The Gateaux derivative wrt to NP is zero $\Rightarrow$ Neyman-orthogonal score

Discussion partially linear Double ML

Advantages:

- Goes beyond linear models
- Allows to plug-in the supervised ML toolbox for nuisance parameter estimation
- Works for binary and continuous treatments

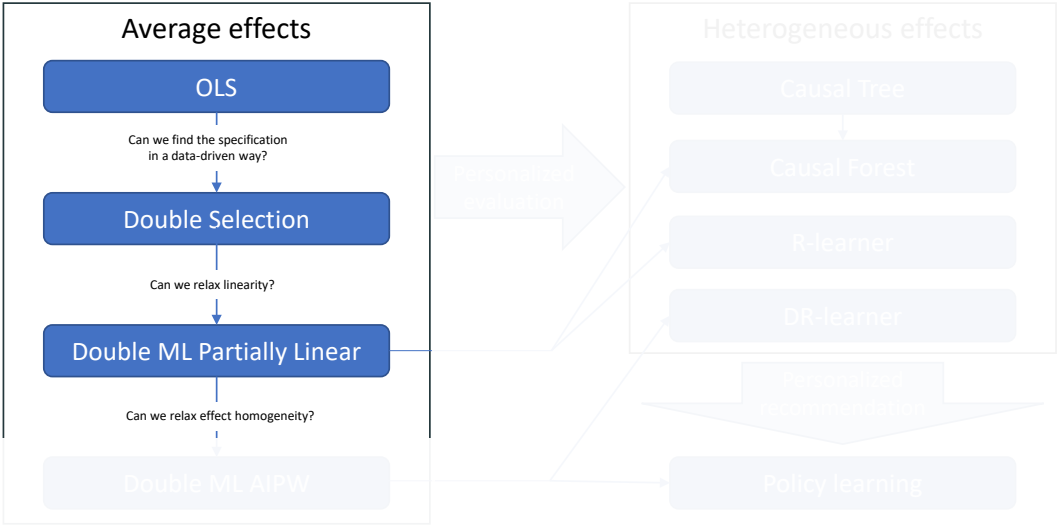
Disadvantages:

- Effect homogeneity often implausible

Simulation notebook: Partially linear Double ML

Application notebook: Double Selection and Partially Linear Double ML (part 2)

Next week


Average effects
OLS
Can we find the specification in a data-driven way?
Double Selection
Can we relax linearity?
Double ML Partially Linear
Can we relax effect homogeneity?
Double ML AIPW
Personalized evaluation
Heterogeneous effects
Causal Tree
Causal Forest
R-learner
DR-learner
Personalized recommendation
Policy learning

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

$\Rightarrow$ separate identification and estimation