Double machine learning for sample selection models

Michela Bia*, Martin Huber**, and Lukáš Lafférs+

*Luxembourg Institute of Socio-Economic Research and University of Luxembourg
**University of Fribourg, Dept. of Economics and
Center for Econometrics and Business Analytics, St. Petersburg State University
+Matej Bel University, Dept. of Mathematics

CBE seminar, NHH, October 19, 2021

+(supported by VEGA-1/0692/20 and APVV-17-0329)

This paper

Estimating **ATEs** with **outcome attrition/sample selection**

based on

double machine learning

under

selection on observables or **instrumental variable** assumptions

Introduction

Treatment evaluation under sample selection

- Examples:
  - Returns to education: wages are only observed for working individuals.
  - Effect of educational interventions (like vouchers for private schools) on college admissions tests: students may non-randomly abstain from the test.
- Typically assumed: selection on observables, see e.g. Imbens (2004).
- Double machine learning (DML, see Chernozhukov et al. 2018) controls for crucial confounders among potentially many covariates in a data-driven way by machine learning.

- Treatment is not random.

- Treatment is not random.
  Observed covariates make the treatment ”as good as random”.

- Treatment is not random.
  Observed covariates make the treatment ”as good as random”.
- Selection is not random.

- Treatment is not random.
  Observed covariates make the treatment ”as good as random”.
- Selection is not random.
  (a) Observed covariates make the selection ”as good as random”.

  or
  (b) There is an instrument for selection.

- Treatment is not random.
  Observed covariates make the treatment ”as good as random”.
- Selection is not random.
  (a) Observed covariates make the selection ”as good as random”.

  or

  (b) There is an instrument for selection.
- Large number of covariates?

- Treatment is not random.
  Observed covariates make the treatment ”as good as random”.
- Selection is not random.
  (a) Observed covariates make the selection ”as good as random”.

  or
  (b) There is an instrument for selection.
- Large number of covariates?
  We make use of ”double machine learning” framework.

Contribution

- DML for discrete treatments under outcome attrition.

Contribution

- DML for discrete treatments under outcome attrition.
- **Static confounding**:
  - selection-on-observables assumption for the treatment
  - selection-on-observables or instrumental variable (IV) assumptions for outcome attrition

Contribution

- DML for discrete treatments under outcome attrition.
- **Static confounding**:
  - selection-on-observables assumption for the treatment
  - selection-on-observables or instrumental variable (IV) assumptions for outcome attrition
- **Dynamic confounding**:
  - sequential selection-on-observables assumption for the treatment and attrition
    (treatment may affect confounders of attrition/outcome)

Contribution

- DML for discrete treatments under outcome attrition.
- **Static confounding**:
  - selection-on-observables assumption for the treatment
  - selection-on-observables or instrumental variable (IV) assumptions for outcome attrition
- **Dynamic confounding**:
  - sequential selection-on-observables assumption for the treatment and attrition
    (treatment may affect confounders of attrition/outcome)
- We derive doubly robust and efficient score functions (see Robins et al. 1994) for treatment evaluation and show that they satisfy the conditions of DML framework.

Contribution

- DML for discrete treatments under outcome attrition.
- **Static confounding**:
  - selection-on-observables assumption for the treatment
  - selection-on-observables or instrumental variable (IV) assumptions for outcome attrition
- **Dynamic confounding**:
  - sequential selection-on-observables assumption for the treatment and attrition
    (treatment may affect confounders of attrition/outcome)
- We derive doubly robust and efficient score functions (see Robins et al. 1994) for treatment evaluation and show that they satisfy the conditions of DML framework.
- $\rightarrow \sqrt{n}$-consistency normality of treatment effect estimation when using machine learners for (first-step) estimation of outcome, selection, and treatment models that converge with rate $n^{-\frac{1}{4}}$.

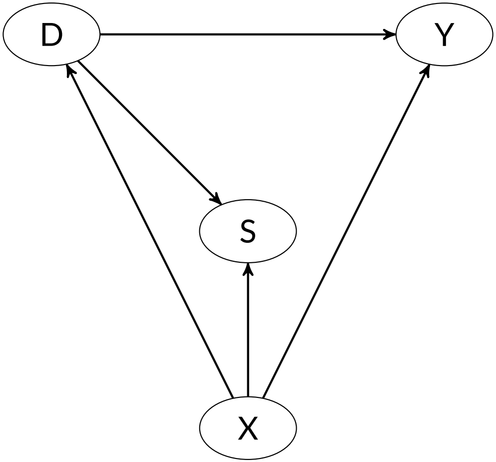
Previous literature Literature

Notation

- D: Treatment.
- Y: Outcome.
- S: Selection indicator.
- X: Covariates.
- $Y(d)$: (Potential) outcome under treatment $d \in \{0, 1, ..., Q\}$.

Identification (MAR)

Identification under MAR (causal graphs):



Identification under MAR

Assumption 1 (conditional independence of the treatment):
$Y(d) \perp D | X = x$

Assumption 2 (conditional independence of selection):
$Y \perp S | D = d, X = x$

Assumption 3 (common support):
(a) $\Pr(D = d | X = x) > 0$ and (b) $\Pr(S = 1 | D = d, X = x) > 0$

Identification under MAR (DR)

- Identification based on the efficient influence function:

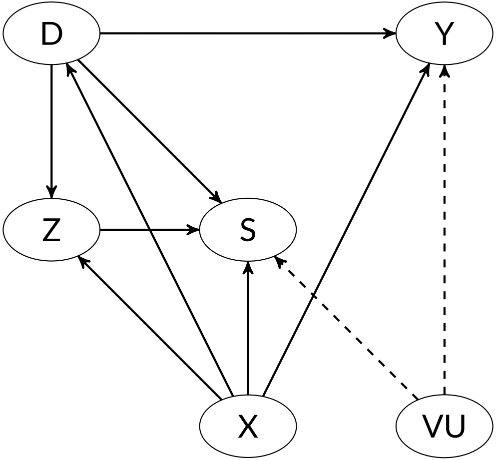
$$
\begin{aligned}
E[Y(d)] &= E\left[\psi_d\right], \text{ where} \\
\psi_d &= \frac{I\{D=d\} \cdot S \cdot [Y - \mu(d,1,X)]}{p_d(X) \cdot \pi(d,X)} + \mu(d,1,X). \qquad (1)
\end{aligned}
$$

where nuisance parameters:

- $\mu(D,S,X) = E[Y|D,S,X]$
- $p^d(X) = \Pr(D=d|X)$
- $\pi(D,X) = \Pr(S=1|D,X)$

Identification (based on IV)

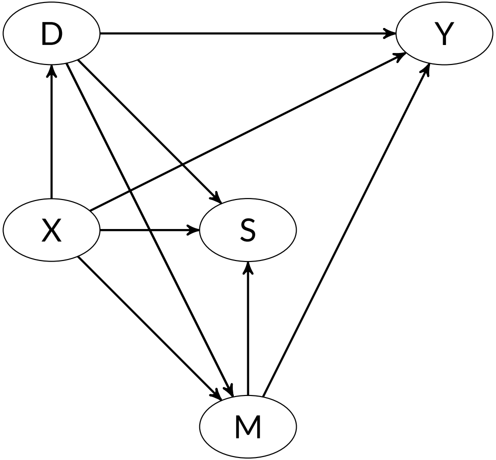
Identification based on IV (causal graphs):



Details

Identification (dynamic confounding)

Sequential conditional independence (graph):

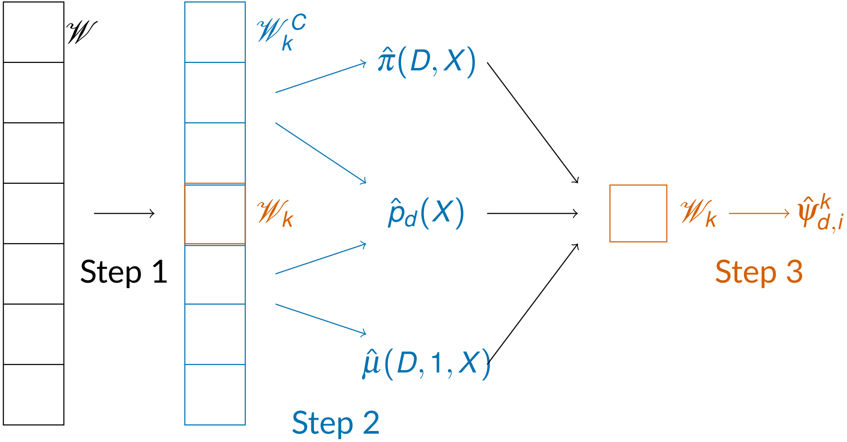


Details

Double machine learning (1)

- X are high-dimensional, the nuisance parameters μ, p_d, π can be estimated with ML algorithms
- ML gives **biased** estimations due to bias-variance trade-off (regularization bias).
- Treatment effect estimation based sample analogs of efficient score functions is quite **robust to regularization bias**
- Neyman-orthogonality - ψ_d is locally insensitive to mild deviations of μ, p_d, π from the true functions μ_0, p_{d0}, π_0

Double machine learning (2)



$$\hat{\psi}_d = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$$

Step 4

Algorithm 1

Regularity conditions

<u>Assumption 10 (regularity conditions and quality of plug-in parameter estimates):</u>

Details

- Satisfied if each nuisance estimator converges at least with rate $n^{-\frac{1}{4}}$ to its true value.
- Can be achieved by common machine learning algorithms like lasso, random forests, neural nets, and boosting.

Root-n consistency

$\Rightarrow$ Treatment effect estimation is $\sqrt{n}$-consistent and asymptotically normal.

- Asymptotic variance is not affected by machine learning.

Theorem 1

Under Assumptions 1-3 and 10, it holds for estimating $\psi_{d0} = E[Y(d)]$ based on Algorithm 1:

$\sqrt{n}\left(\hat{\psi}_d - \psi_{d0}\right) \rightarrow N(0, \sigma^2_{\psi_d})$, where $\sigma^2_{\psi_d} = E[(\psi_d - \psi_{d0})^2]$.

Simulation study

Data generating process:

$$
\begin{aligned}
Y &= D + X'\beta + U \text{ with } Y \text{ being observed if } S = 1, \\
S &= I\{D + \gamma Z + X'\beta + V > 0\}, \quad D = I\{X'\beta + W > 0\}, \\
X &\sim N(0, \sigma_X^2), \quad Z \sim N(0,1), \quad (U,V) \sim N(0, \sigma_{U,V}^2), \quad W \sim N(0,1).
\end{aligned}
$$

Details on simulation

Application: Job Corps experimental study

- Job Corps offers **vocational** training and **academic** classroom instruction for **disadvantaged individuals aged 16 to 24**
- Currently about 50,000 participants every year.
- Sample comes from the Job Corps **experimental study** conducted in **mid-90's**, see Schochet et all (2008): 11313 young individuals with completed interviews 4 years after randomization (6828 assigned to Job Corps, 4485 randomized out).
- Outcome Y is **hourly wage** in last week of first year or four years after randomization, observed conditional on employment S.
- Treatment D is participation in academic or vocational **training** in the first year after randomization among those randomized in.

Application

- Focussing on female subsample randomized into Job Corps.
- **Hundreds of baseline covariates** X (socioeconomic vars, labor market history, crime, health...).
- Instrument Z: number of young children in the household at baseline.
- $\rightarrow$ DML IV (Theorem 3) to assess ATE on hourly wage at the end of first year.
- **Hundreds of intermediate covariates** M measured after one year.
- $\rightarrow$ DML under sequential selection on observables (Theorem 4) to assess ATE on hourly wage after four years.
- Random forests for nuisance parameter estimation and 3-fold cross-validation.

Application

Evaluation sample:

Table: Treatment distribution

treatment	observations
randomized out of JC	1698
controls (no training)	200
academic training	830
vocational training	843

Application

Results:

Table: ATE estimates

$D=1$	$D=0$	ATE	se	p-value
		Theorem 1 (MAR)		
academic	no training	-0.170	0.253	0.501
vocational	no training	-0.519	0.405	0.199
		Theorem 3 (IV)		
academic	no training	-0.192	0.174	0.705
vocational	no training	-0.537	0.404	0.199
		Theorem 4 (sequential)		
academic	no training	0.170	0.117	0.147
vocational	no training	0.442	0.096	0.000

Conclusion

- Evaluation of **average treatment effects** in the presence of **sample selection or outcome attrition** based on **double machine learning**.
- Proposition of doubly robust and Neyman-orhtogonal estimators that are $\sqrt{n}$-consistent and asymptotically normal under specific regularity conditions.
- Simulation study and application to Job Corps program.
- In `causalweight` package for R by Bodory and Huber (2018).

Thank you for your attention.

Literature

- Selection-on-observables/missing at random (MAR) assumption for outcome attrition: Rubin (1976), Little and Rubin (1987), Fitzgerald et al. (1998), Wooldridge (2002, 2007)...
- Doubly robust (DR) estimation under MAR: Robins et al. (1994, 1995), Bang and Robins (2005) - can satisfy DML framework, but treatment selection not considered.
- Negi (2020): weighted M-estimator under double selection (static, MAR) satisfying DR (consistent under parametric misspecification of either the conditional outcome model or the treatment and selection models), but unclear whether DML conditions are met.
- Nonignorable non-response models for outcome attrition (using parametric assumptions or IV): Heckman (1976, 1979), Hausman and Wise (1979), Little (1995), Das et al. (2003), Newey (2007)....
- Double selection (static, MAR or IV): Huber (2012, 2014) using inverse probability weighting (not DR). Back

Identification based on IV (assumptions)

Assumption 4 (Instrument for selection):

(a) $E[Z \cdot S|D,X] \neq 0$,
$Y(d,z) = Y(d)$, and
$Y \perp Z | D = d, X = x$

(b) $S = I\{V \leq \chi(D,X,Z)\}$,

(c) $V \perp (D,Z)|X$.

Assumption 5 (common support):

$\Pr(D = d|X = x, \Pi = \pi) > 0$,

where

- $\Pi = \pi(D,X,Z) = \Pr(S = 1|D,X,Z)$.

Identification based on IV (assumptions)

Assumption 6 (conditional effect homogeneity):

$E[Y(d) - Y(d')|S = 1, X = x, V = v] = E[Y(d) - Y(d')|X = x, V = v]$

- Effect homogeneity is satisfied if unobservables in the outcome equation are additive separable.

Assumption 7 (common support):

$\pi(d, x, z) > 0$

Identification based on IV (DR):

- Under Assumptions 1, 4, and 5:

$$E[Y(d)|S=1] = E\left[\phi_{d,S=1}|S=1\right], \text{ where}$$

$$\phi_{d,S=1} = \frac{I\{D=d\}\cdot[Y-\mu(d,1,X,\Pi)]}{p_d(X,\Pi)} + \mu(d,1,X,\Pi).$$

- Under Assumptions 1, 4, 5, 6, and 7:

$$\Delta = E\left[\phi_d - \phi_{d'}\right], \text{ where}$$

$$\phi_d = \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X,\Pi)]}{p_d(X,\Pi)\cdot\pi(d,X,Z)} + \mu(d,1,X,\Pi). \tag{2}$$

- $p_d(X,\Pi) = \Pr(D=d|X,\Pi)$
- $\mu(D,S,X,\Pi) = E[Y|D,S,X,\pi(D,X,Z)]$

Identification (dynamic confounding)

Assumption 8 (conditional independence of selection):
$Y \perp S | D = d, X = x, M = m$
Assumption 9 (common support):
(a) $\Pr(D = d | X = x) > 0$ and (b) $\Pr(S = 1 | D = d, X = x, M = m) > 0$

- M – post-treatment covariates.

Identification (dynamic confounding)

- Under Assumptions 1, 8, and 9:

$$
\begin{aligned}
E[Y(d)] &= E\left[\theta_d\right], \text{ where} \\
\theta_d &= \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X,M)]}{p_d(X)\cdot\pi(d,X,M)} \\
&+ \frac{I\{D=d\}\cdot[\mu(d,1,X,M)-\nu(d,1,X)]}{p_d(X)} + \nu(d,1,X).
\end{aligned}
\tag{3}
$$

- $\pi(D,X,M) = \Pr(S=1|D,X,M)$
- $\mu(d,1,X,M) = E[Y|D=d,S=1,X,M]$
- $\nu(d,1,X) = E[E[Y|D=d,S=1,X,M]|D=d,X]$

Double machine learning

- Risk of overfitting bias when estimating nuisance terms μ, p^d, π in the same sample as the treatment effect.

$\Rightarrow$ Cross-fitting: randomly split data to

(i) estimate the model parameters of nuisance terms in one subsample and
(ii) predict nuisance terms/estimate treatment effects in another subsample.

- Subsamples are like independently drawn samples.
- Switch roles of subsamples to avoid efficiency loss.

Double machine learning

Algorithm 1: Estimation of $E[Y(d)]$ based on equation (1)

- Let $\mathscr{W} = \{W_i | 1 \leq i \leq n\}$ with $W_i = (Y_i \cdot S_i, D_i, S_i, X_i)$ for all i denote the set of observations in an i.i.d. sample of size n.

1. Split $\mathscr{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathscr{W}_k$ the set of observations in the sample and $\mathscr{W}_k^C$ the complement set of all observations not in k.
2. For each k, use $\mathscr{W}_k^C$ to estimate the model parameters of the plug-ins $\mu(D,S=1,X)$, $p_d(X)$, $\pi(D,X)$ in order to predict these plug-ins in $\mathscr{W}_k$, where the predictions are denoted by $\hat{\mu}^k(D,1,X)$, $\hat{p}_d^k(X)$, and $\hat{\pi}^k(D,X)$.
3. For each k, obtain an estimate of the score function (see ψ_d in (1)) for each observation i in $\mathscr{W}_k$, denoted by $\hat{\psi}_{d,i}^k$:

$$\hat{\psi}_{d,i}^k = \frac{I\{D_i = d\} \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i)]}{\hat{p}_d^k(X_i) \cdot \hat{\pi}^k(d,X_i)} + \hat{\mu}^k(d,1,X_i). \tag{4}$$

4. Average the estimated scores $\hat{\psi}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\Psi_{d0} = E[Y(d)]$ in the total sample, denoted by $\hat{\Psi}_d = 1/n \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$.

Double machine learning (2)

Regularity conditions and root-n consistency:

Assumption 10 (regularity conditions and quality of plug-in parameter estimates):
For all probability laws $P \in \mathscr{P}$, where $\mathscr{P}$ is the set of all possible probability laws the following conditions hold for the random vector (Y, D, S, X) for $d \in \{0, 1, ..., Q\}$:

(a) $\|Y\|_q \leq C$, $\left\|E[Y^2|D=d, S=1, X]\right\|_\infty \leq C^2$,

(b) $\Pr(\varepsilon \leq p_{d0}(X) \leq 1-\varepsilon) = 1$, $\Pr(\varepsilon \leq \pi_0(d, X)) = 1$,

(c) $\|Y - \mu_0(d, 1, X)\|_2 = E\left[(Y - \mu_0(d, 1, X))^2\right]^{\frac{1}{2}} \geq c$

(d) Given a random subset I of $[n]$ of size $n_k = n/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^C})$ satisfies the following conditions. With P-probability no less than $1 - \Delta_n$:

$$\|\hat{\eta}_0 - \eta_0\|_q \leq C, \quad \|\hat{\eta}_0 - \eta_0\|_2 \leq \delta_n,$$
$$\|\hat{p}_{d0}(X) - 1/2\|_\infty \leq 1/2 - \varepsilon, \quad \|\hat{\pi}_0(D, X) - 1/2\|_\infty \leq 1/2 - \varepsilon,$$
$$\|\hat{\mu}_0(D, S, X) - \mu_0(D, S, X)\|_2 \times \|\hat{p}_{d0}(X) - p_0(X)\|_2 \leq \delta_n n^{-1/2},$$
$$\|\hat{\mu}_0(D, S, X) - \mu_0(D, S, X)\|_2 \times \|\hat{\pi}_0(D, X) - \pi_0(D, X)\|_2 \leq \delta_n n^{-1/2}.$$

Simulation study

Simulation design MAR:

- Dimension of X: $p = 100$, number of simulations: 1000.
- ith element in the coefficient vector β is set to $0.4/i^2$ for $i = 1, \ldots, p$.
- σ_X^2 is defined based on setting the covariance of the ith and jth covariate in X to $0.5^{|i-j|}$.
- Sample sizes: $n = 2,000$ and $n = 8,000$.
- $\gamma = 0$ and $\sigma_{U,V}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.
- DML based on Theorem 1 (henceforth DML MAR) and Theorem 2 (DML IV - uses instrument Z despite satisfaction of MAR).
- Estimation based on 3-fold cross-fitting with nuisance terms obtained by lasso regression.

Simulation study

Results MAR:

Table: Simulation results under MAR

	true	bias	sd	RMSE	meanSE	coverage
n=2000						
DML MAR	1.000	0.003	0.060	0.060	0.063	0.939
DML IV	1.000	0.003	0.060	0.060	0.063	0.939
n=8000						
DML MAR	1.000	0.012	0.031	0.033	0.034	0.934
DML IV	1.000	0.012	0.031	0.033	0.034	0.939

Simulation study

Simulation design IV:

- In a second simulation design, we set $\gamma = 1$ and $\sigma^2_{U,V} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, such that selection is nonignorable due to the correlation of U and V.
- DML MAR is no longer unbiased, while the bias of DML IV appears to approach zero as the sample size increases, at the price of somewhat higher standard deviation than DML MAR.

Simulation study

Results IV:

Table: Simulation results under nonignorable selection

	true	bias	sd	RMSE	meanSE	coverage
n=2000						
DML MAR	1.000	-0.120	0.055	0.132	0.052	0.374
DML IV	1.000	-0.020	0.071	0.074	0.065	0.907
n=8000						
DML MAR	1.000	-0.116	0.028	0.119	0.027	0.009
DML IV	1.000	0.006	0.040	0.040	0.036	0.915

Go back