Causal Effects Estimation and Machine Learning

Lukáš Lafférs

Matej Bel University, Dept. of Mathematics

Habilitation lecture

Motivation

Job-seeker went through a training/course. Did it help?

We know a lot about these job-seekers (say 300 variables).

But sample size is small.

Motivation (cont'd)

More information is desirable. Traditional models are not feasible.

It helps with

- statistical precision - reduces uncertainty
- identification - treated and non-treated units are more comparable

Also, we wish to have flexible model specification.

Can ML algorithms help??

This presentation

Indeed, ML algorithms can help.

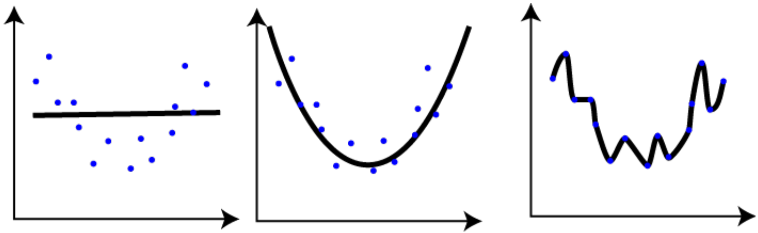
Introduction to Double Machine Learning framework

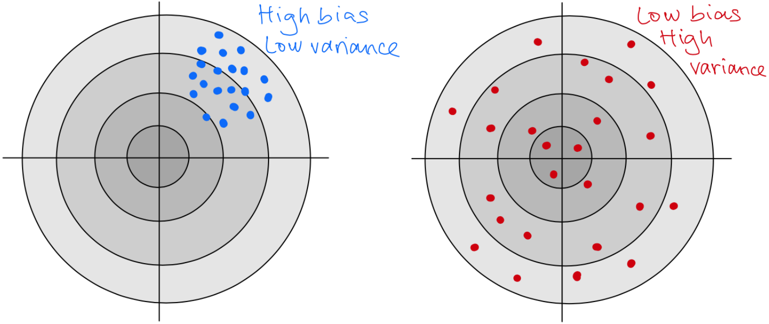
Three extensions

Machine learning and causality

ML is (mostly) about prediction.

While ML predicts well, we are often interested in a **certain variable of interest**.




High bias
Low variance
Low bias
High
variance

Can we make use of the great predictive capabilities of ML algorithms for improving the estimation of parameters of interest?

This is an area of active research: **DOUBLE MACHINE LEARNING**

Seminal paper

Double machine learning

Victor, Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. : ”Double/debiased machine learning for treatment and structural parameters.” The Econometrics Journal 21.1 (2018): C1-C68.

Double Machine Learning framework

Example: Consider the following partially linear model. θ is the parameter of interest. $g(X)$ and $m(X)$ are some flexible functions, not of interest

$$
\begin{aligned}
Y &= \theta D + g(X) + U, \qquad && E[U|D,X] = 0 \\
D &= m(X) + V, \qquad && E[V|X] = 0
\end{aligned}
$$

Split the data into two parts

- Use the first one to get $\hat{g}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{\theta}_1$ from regressing $Y - \hat{g}(X)$ on D

Naive approach: $\hat{\theta}_1$

How does this naive estimator $\hat{\theta}_1$ behave?

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i U_i}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i(g(X_i) - \hat{g}(X_i))}_{\text{In general divergent.}}$$

So it leads to a regularization bias.

Alternative approach: $\hat{\theta}_2$

Instead of $\hat{\theta}_1$, we will do something else:

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{\theta}_2$ by regressing $Y - \hat{g}(X)$ on $D - \hat{m}(X)$

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

- Regularization bias : ML for $\hat{g}$ and $\hat{m}$ are allowed to converge "slowly"
- Overfitting bias: Sample splitting takes care of this.

$\hat{\theta}_1$ is based on moment condition
$\psi_1 = D(Y - g(X) - \theta D)$

$\hat{\theta}_2$ is based on moment condition
$\psi_2 = (D - m(X)) \cdot (Y - g(X) - \theta D)$

What makes ψ_2 different from ψ_1 ???

Regularization bias vanishes under mild conditions.

In other words, ψ_2 is locally insensitive to some mild perturbations of $\hat{m}, \hat{g}$ around m, g.

This local insensitiveness has a name: **Neyman-orthogonality**.

$$E[\psi(W; \theta_0, \eta_0)] = 0.$$

- ψ is a moment condition
- θ is the parameter of interest (target parameter)
- $\eta = (m, g)$ is the nuisance parameter vector
- $W = (Y, D, X)$ denotes data

In a small neighborhood of η_0, ψ does not change much:

$$\frac{\partial}{\partial r} E[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]\Big|_{r=0} = 0$$

Neyman-orthogonality

Simple calculations show that

- ψ_1 is not locally insensitive to bias in η Details
- ψ_2 is locally insensitive to bias in η Details

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

Overfitting bias may arise from the fact that the same data is used for both estimation of nuisance functions and target parameter.

We can **split the data**. As we already did.

$\rightarrow$ But then we loose many observations.

How to fix this? **Swap the roles** of the two data parts and then average across them!

Details

DML wrap-up (1)

We saw : $\hat{\theta}_1$ and $\hat{\theta}_2$.

Based on: ψ_1 and ψ_2.

While ψ_1 was locally sensitive to some small changes in the η, the other ψ_2 was not.

This allows us to get rid of the regularization bias.

Sample-splitting removes the overfitting bias.

DML wrap-up (2)

- Estimator $\hat{\theta}$ based on Neyman-orthogonal moment function ψ
- Apply sample splitting
- Nuisance parameter estimators m and g are ”good enough”
  (e.g. converge at rate at least $n^{-1/4}$)

Theorem 1 in Chernozhukov et al. 2018:

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma^2)$$

Asymptotically normally distributed estimator that is $\sqrt{n}$ consistent.

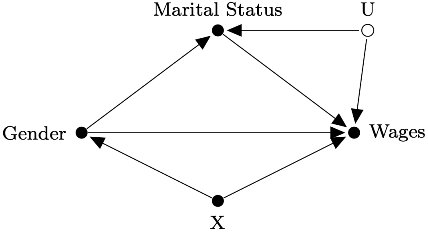
DML final wrap-up

DML provides a framework for developing estimators that:

- can handle high-dimensional data
- are flexible
- make use of predictive powers of ML

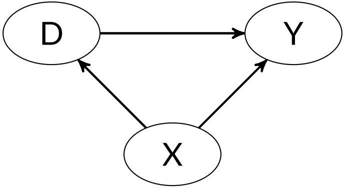
This addresses all the points in the motivation!

Limitations - ”Kitchen sink” regression



Hünermund, Beyers and Caspi (2023)

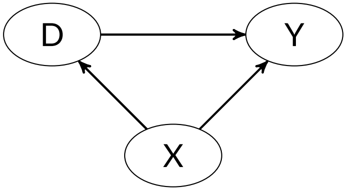
DML and policy evaluation



Notation:

- $Y(d)$: (Potential) outcome as function of treatment d
- Y - outcome
- D - treatment
- X - covariates

DML and policy evaluation



Object of interest:

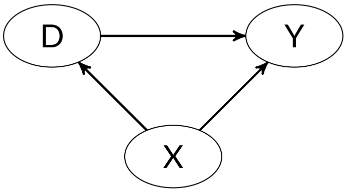
$$\Delta = E[Y(1) - Y(0)]$$

Indentifying assumptions:

1) Conditional independence of D:
$Y(d) \perp D \mid X$

2) Common support:
$\Pr(D = d | X = x) > 0$

DML and policy evaluation



Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot[Y_2-\mu(d,X)]}{p(X)} + \mu(d,X) - \theta_0.$$

$$E\Big[\psi(W;\theta_0,\eta)\Big] = E\Big[Y(d)\Big] - \theta_0 = 0$$

Data: $W = (Y, D, X)$

Nuisance functions: $\eta = (p, \mu)$

- $p(X) \equiv \Pr(D = d|X)$
- $\mu(D,X) \equiv E[Y|D,X]$

Bang, Heejung, and James M. Robins. "Doubly robust estimation in missing data and causal inference models." Biometrics 61.4 (2005): 962-973.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. The Econometrics Journal, 25(3), 602-627.

DML applications

There are **many**:

Double/debiased machine learning for treatment and structural parameters [HTML] oup.com

V Chernozhukov, D Chetverikov, M Demirer, E Duflo… - 2018 - academic.oup.com

… To estimate η 0 , we consider the use of statistical or **machine learning** (ML) methods, which are … We call the resulting set of methods **double** or **debiased** ML (DML). We verify that DML …

☆ Save 99 Cite Cited by 2765 Related articles All 28 versions

Most read

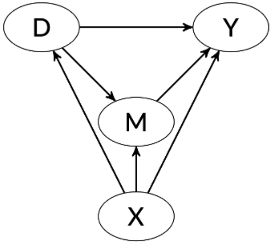
Double/debiased machine learning for treatment and structural parameters

DML extensions

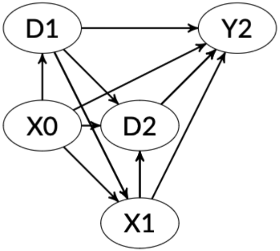
- mediation analysis (H. Farbmacher, M. Huber, H. Langen, L. Lafférs, M. Spindler)
- dynamic treatment effects (H. Bodory, M. Huber, L. Lafférs)
- sample selection models (M. Bia, M. Huber, L. Lafférs)

DML extensions

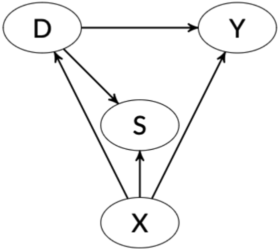
Mediation analysis



Dynamic treatment effects



Sample selection models

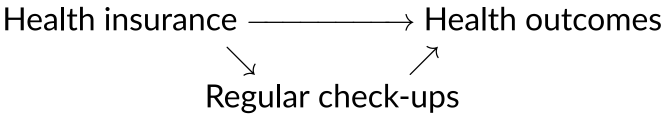


First extension

DML and mediation analysis

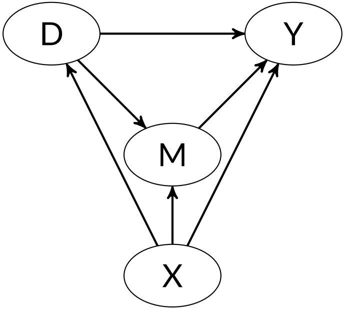
Helmut Farbmacher, Martin Huber, Lukáš Lafférs, Henrika Langen and Martin Spindler: Causal mediation analysis with double machine learning (Econometrics Journal, 2022, 25 (2), 277—300)

Example



Details Examples Mediation analysis

DML and mediation analysis



Objects of interest:

Indirect effect: $E[Y(d, M(1)) - Y(d, M(0))]$
Direct effect: $E[Y(1, M(d)) - Y(0, M(d))]$

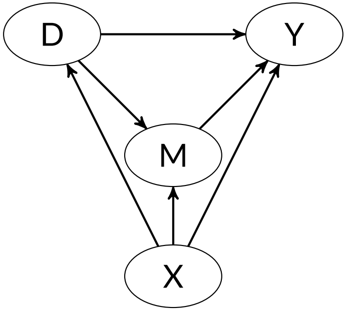
Indentifying assumptions:

1) Conditional independence of D

2) Conditional independence of M

3) Common support

DML and mediation analysis



Moment function:

$$\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}(1-p_d(M,X))}{p_{dm}(M,X)\cdot 1-p_d(X)}\cdot[Y-\mu(d,M,X)] \\
&+ \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X)-\omega(1-d,X)\Big] \\
&+ E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y(d,M(1-d))\Big]-\theta_0=0
\end{aligned}$$

Data: $W=(Y,D,M,X)$

Nuisance functions: $\eta=(p_d,p_{dm},\mu,\omega)$

- $p_d(X)=Pr(D=d|X)$
- $p_{dm}(M,X)=Pr(D=d|M,X)$
- $\mu(D,M,X)=E(Y|D,M,X)$
- $\omega(1-d,X)=E[\mu(d,M,X)|D=1-d,X]$

Application

Results:

		direct		indirect	
	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
	Modified score using Bayes' rule				
effect	-0.05	-0.07	-0.05	0.00	0.02
se	0.03	0.03	0.03	0.01	0.01
p-value	0.10	0.03	0.10	0.89	0.07

- Health insurance coverage appears to moderately improve general health in the short run among young adults in the U.S. through mechanisms other than routine checkups.

Details

Second extension

DML and dynamic treatment effects

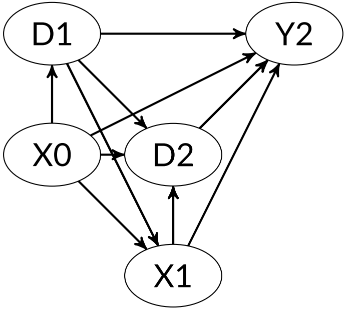
Hugo Bodory, Martin Huber and Lukáš Lafférs: Evaluating (weighted) dynamic treatment effects by double machine learning (The Econometrics Journal 25.3 (2022): 628–648

Example

Academic/vocational Trainings $\longrightarrow$ Employment

Details

DML and dynamic treatment effects



Objects of interest:

$$E[Y(\underline{d}_2)] - E[Y(\underline{d}_2^*)]$$

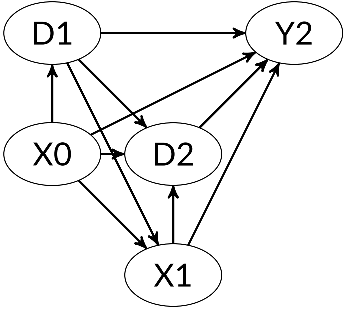
Indentifying assumptions:

1) Conditional ind. of the first treatment

2) Conditional ind. of the second treatment

3) Common support

DML and dynamic treatment effects



Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D_1=d_1\}\cdot I\{D_2=d_2\}\cdot[Y_2-\mu^{Y_2}(\underline{d}_2,\underline{X}_1)]}{p^{d_1}(X_0)\cdot p^{d_2}(d_1,\underline{X}_1)}$$

$$+ \frac{I\{D_1=d_1\}\cdot[\mu^{Y_2}(\underline{d}_2,\underline{X}_1)-\nu^{Y_2}(\underline{d}_2,X_0)]}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2,X_0)-\theta_0.$$

$$E\left[\psi(W;\theta_0,\eta)\right] = E\left[Y_2(\underline{d}_2)\right]-\theta_0 = 0$$

Data: $W = (Y_2, D_1, D_2, X_0, X_1)$

Nuisance functions: $\eta = (p^{d_1}, p^{d_2}, \mu^{Y_2}, \nu^{Y_2})$

- $p^{d_1}(X_0) \equiv \Pr(D_1 = d_1|X_0)$
- $p^{d_2}(D_1,\underline{X}_1) \equiv \Pr(D_2 = d_2|D_1,\underline{X}_1)$
- $\mu^{Y_2}(\underline{D}_2,\underline{X}_1) \equiv E[Y_2|\underline{D}_2, X_0, X_1]$
- $\nu^{Y_2}(\underline{D}_2,X_0) \equiv E[E[Y_2|\underline{D}_2, X_0, X_1]|D_1, X_0]$,

DML and dynamic treatment effects: Application

Results (outcome: employment after 4 years):

3 = vocational 2=academic

10% ↑ employment after 4 years

$\underline{d}_2$	$\underline{d}_2^*$	$\hat{E}[Y_2(\underline{d}_2^*)\|S=1]$	$\hat{\Delta}(\underline{d}_2, \underline{d}_2^*, S=1)$	SE	p-value	observations	trimmed
33	22	0.76	0.1	0.06	0.11	3783	507
33	21	0.82	0.05	0.03	0.07	3783	43
33	11	0.81	0.08	0.03	0.02	2346	22

1= no training

Details

Third extension

DML and sample selection models

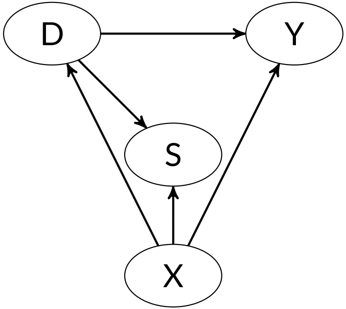
Michela Bia, Martin Huber and Lukáš Lafférs: Double machine learning for sample selection models. Journal of Business & Economic Statistics, (forthcoming)

Example

Academic/vocational Training $\longrightarrow$ Wages

Details

DML and sample selection models



Object of interest:

$$E[Y(d)] - E[Y(d^*)]$$

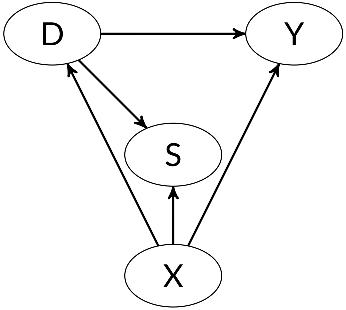
Indentifying assumptions

1) Conditional independence of the treatment:

2) Conditional independence of selection

3) Common support

DML and sample selection models



Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)} + \mu(d,1,X) - \theta_0.$$

$$E\left[\psi(W;\theta_0,\eta)\right] = E\left[Y(d)\right] - \theta_0 = 0$$

Data: $W = (Y.S, S, D, X)$

Nuisance functions: $\eta = (p^d, \pi, \mu)$

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

DML and sample selection models: Application

$D = 1$	$D = 0$	ATE	standard error	p-value
		Theorem 1 (MAR)		
academic	no training	-0.683	1.073	0.524
vocational	no training	0.611	0.629	0.331
		Theorem 3 (IV)		
academic	no training	-0.631	1.052	0.549
vocational	no training	0.586	0.645	0.364
		Theorem 4 (sequential)		
academic	no training	0.149	0.199	0.454
vocational	no training	0.567	0.208	0.007

We observe small longer-term wage gains in terms of hourly wage.

Details

Recapitulation

DML is a useful framework for estimation under high-dimensional setting.

It can automatically select among many covariates and avoid

regularization bias (via Neyman-orthogonality) and
overfitting bias (via cross-fitting) and

provide root-n consistent and asymptotically normal estimator.

I have shown three extensions of DML that appear to be empirically relevant and useful.

Implemented in `causalweight` R package (Bodory and Huber 2018)

Thank you.

References

- Double machine learning framework: Chernozhukov, Victor, et al. ”Double/debiased machine learning for treatment and structural parameters.” The Econometrics Journal 21.1 (2018): C1-C68.
- DoubleML package in R `https://cran.r-project.org/web/packages/DoubleML/DoubleML.pdf`
- Bach, Philipp, et al. ”DoubleML–An Object-Oriented Implementation of Double Machine Learning in R.” arXiv preprint arXiv:2103.09603 (2021).
- Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. The Econometrics Journal, 25(3), 602-627.
- Bang, Heejung, and James M. Robins. ”Doubly robust estimation in missing data and causal inference models.” Biometrics 61.4 (2005): 962-973.
- Hünermund, P., Louw, B., and Caspi, I. (2023). Double machine learning and automated confounder selection: A cautionary tale. Journal of Causal Inference, 11(1), 20220078.
- Farbmacher, Helmut, et al. ”Causal mediation analysis with double machine learning.” The Econometrics Journal 25.2 (2022): 277-300.
- Bodory H., Huber M. and Lafférs L. ”Evaluating (weighted) dynamic treatment effects by double machine learning.” The Econometrics Journal 25.3 (2022): 628—648.
- Bia, M., Huber, M., and Lafférs, L. (2023). Double machine learning for sample selection models. Journal of Business & Economic Statistics, 1-12.
- Bodory, Hugo, and Martin Huber. ”The causalweight package for causal inference in R.” (2018).

Mediation Example - details

- health insurance coverage $\rightarrow$ general health (self-reported)
- health insurance coverage $\rightarrow$ regular checkups $\rightarrow$ general health
- X - demographics, family background, education, labor market, household char, mental health, nutrition, physical activity.... (755 control variables, from 2005)

- National Longitudinal Survey of Youth 1997 (NLSY97), a survey by the US Department of Labor (2019) (n $\approx$ 7500)
- most studies find significant effect on a particular type of screening (cancer, stroke...)
- we have younger individuals and short-term effects (2006 $\rightarrow$ 2007 $\rightarrow$ 2008)
- health - "excellent" to "poor", negative ATE $\approx$ improvement

Back

Related to job training evaluation:

- Direct earning effect of Job Corps training programme using work experience as mediator (Flores and Flores-Lagunes (2009))
- Effect of Perry Preschool Program on healthy behaviour mediated by personality traits (Conti, Heckman and Pinto (2016))
- What is the effect of more rigorous caseworkers in the counselling process on the employment mediated by placement into labor market programme (Huber, Lechner and Mellace (2017))

Related to education and wages:

- How growing up poor affects economic outcomes in adulthood using education as mediator. (Bellani and Bia (2018))
- Wage-gap decomposition (gender, socioeconomic variables, wage) (Huber (2015))
- Effect of education on mortality mediated by cognitive ability (Bijwaard and Jones (2018))

Based on instrumental variables:

- The effect of education on life-satisfaction using income as mediator (Powdthavee, Lekfuangfu and Wooden (2013))
- The effect of education on health mediated by health-behaviour as mediator (Brunello, Fort, Schneeweis and Winter-Ebmer (2016))
- The effect of family composition on the education of the first-born child using family size as mediator (Chen, Chen and Liu (2017))

Back

Dynamic Example - details

- Training $\rightarrow$ Employment (after 4 years)
- X - 1184 variables ($X_0 - 814$, $X_1 - 374$) socio-economic characteristics, pre-treatment education and training, labor market histories, job search activities, welfare receipt, health, crime...

- 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 four years after randomization (6828 assigned to Job Corps, 4485 randomized out).
- Treatment sequences are based on participation in academic or vocational training in the first or second year after randomization among those randomized in.

Back

Sample Selection Example - details

- Training $\rightarrow$ Hourly wage
- Hundreds of baseline covariates X (socioeconomic vars, labor market history, crime, health...).

- 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 - ($n \approx 3600$)
- 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.

Back

Neyman-orthogonality of ψ_2

We will verify that ψ_2 satisfy the Neyman-orthogonality condition, while ψ_1 does not.

Notation

- $\eta = (m, g)$ is the vector of nuisance parameters, $\theta_0 = (m_0, g_0)$ is the true one
- $\eta_r = \eta_0 + r(\eta - \eta_0)$.

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\psi_2(W;\theta_0,\eta_r) &= (D - m_0(X) - r(m(X) - m_0(X))) \cdot (Y - g_0(X) - r(g(X) - g_0(X)) - D\theta_0) \\
&= (D - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) + \\
&\quad - r(D - m_0(X)) \cdot (g(X) - g_0(X)) \\
&\quad - r(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) \\
&\quad + r^2(m(X) - m_0(X)) \cdot (g(X) - g_0(X)) \\
\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)] &= -E[(D - m_0(X)) \cdot (g(X) - g_0(X))] \\
&\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)] \\
&\quad + 2 \cdot r \cdot E[(m(X) - m_0(X)) \cdot (g(X) - g_0(X))] \\
\left.\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)]\right|_{r=0} &= -E[(D - m_0(X)) \cdot (g(X) - g_0(X))] \\
&\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)]
\end{aligned}
$$

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[(D-m_0(X))\cdot(g(X)-g_0(X))] \\
&\quad -E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] \\
&= 0
\end{aligned}
$$

because

$$
\begin{aligned}
E[(D-m_0(X))\cdot(g(x)-g_0(X))] &= E[(g(X)-g_0(X))\cdot\underbrace{E[D-m_0(X)|X]}_{E[V|X]=0}] = 0 \\
E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] &= E[(m(X)-m_0(X))\cdot\underbrace{E[Y-g_0(X)-D\theta_0|X,D]}_{E[U|X,D]=0}] = 0
\end{aligned}
$$

and hence ψ_2 is indeed Neyman-orthogonal.

Back

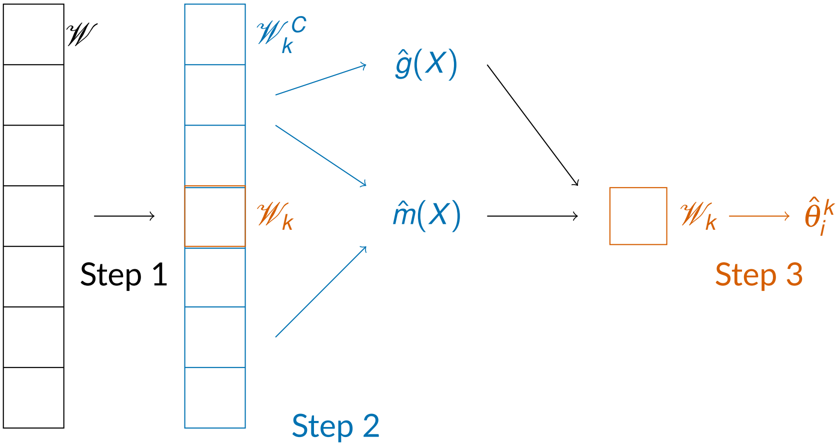
Neyman-orthogonality of ψ_1 ???

$$
\begin{aligned}
\psi_1(W;\theta_0,\eta_r) &= D\cdot(Y-g_0(X)-r(g(X)-g_0(X))-D\theta_0) \\
\frac{\partial}{\partial r}E[\psi_2(W;\theta_0,\eta_r)] &= -E[D\cdot(g(X)-g_0(X))] \\
\frac{\partial}{\partial r}E[\psi(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[D\cdot(g(X)-g_0(X))] \\
&\neq 0
\end{aligned}
$$

There is nothing we could do to use $E[U|X,D]=0$ and $E[V|X]=0$ to make this term equal to zero.

Back

Sample splitting for dealing with overfitting bias



$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\theta}_i^k$$

Step 4

Back