

Kauzálné modely v ekonometrii a metódy strojového učenia

Lukáš Lafférs

Katedra matematiky, Univerzita Mateja Bela

Konferencia slovenských matematikov 2025

Kauzálné modely v **ekonometrii** a metódy strojového učenia

Applied Statistics:

- Biostatistics
- Sociometrics
- Epidemiology
- Survey Research Methods
- **Econometrics**
- Actuarial Statistics
- Reliability statistics
- Environmental statistics
- ...

Kauzálné modely v **ekonometrii** a metódy strojového učenia

Applied Statistics:

- Biostatistics
- Sociometrics
- Epidemiology
- Survey Research Methods
- **Econometrics**
- Actuarial Statistics
- Reliability statistics
- Environmental statistics
- ...

"Econometrics is what Econometricians do." (Arthur Goldberger)

Kauzálné modely v ekonometrii a metódy strojového učenia

Algoritmy strojového učenia sa ukázali ako skvelý nástroj na predikciu!

Kauzálné modely v ekonometrii a metódy strojového učenia

Vedia nám však pomôcť aj s odhadovaním kauzálnych vzťahov??

Kauzalita

Medicína

- Causality - C-word?
- Experiment alebo nie.

Kauzalita

Medicína

- Causality - C-word?
- Experiment alebo nie.

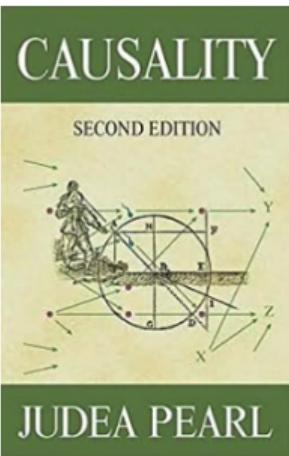
Ekonómia

- Inštrumentálne premenné (Imbens, G. W., and Angrist, J. D. (1994).)
- Dizajn regresnej diskontinuity (Hahn, J., Todd, P., and Van der Klaauw, W. (2001))
- Metóda syntetickej kontrolnej premennej (Abadie, A., and Gardeazabal, J. (2003).)
- Metóda rozdielu v rozdieloch (difference-in-differences) (Card, D., and Krueger, A. (1994).)

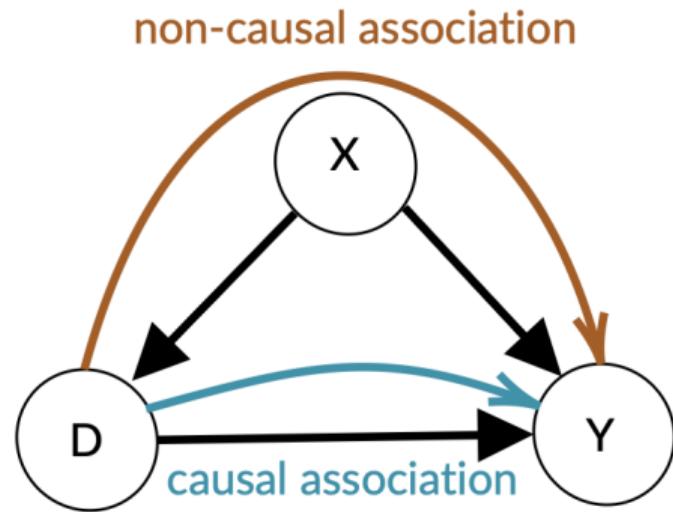
Kauzalita

Informatika/Počítačová veda

- Usmernené acyklické grafy (DAG - Directed Acyclic Graphs)



Judea Pearl (UCLA) a jeho kniha Causality.



Kauzalita

Prečo sa to stane?

Vyžaduje hloboké porozumenie problému.

Náročný problém.

Predikcia

Čo sa stane?

Pokiaľ funguje, nevadí, že nerozumieme prečo.

Principiálne jednoduchšie.

Michal absolvoval rekvalifikačný kurz. Pomohlo to?

Michal absolvoval rekvalifikačný kurz. Pomohlo to?

Mohli by sme ho porovnať s **Jozefom**, ktorý ho neabsolvoval.

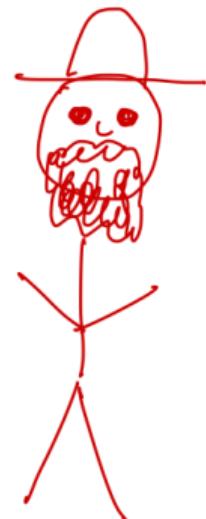
Michal



- 27 rokov
- ženatý
- z Brezna
- stredná škola
- vie anglicky
- zdravý
- vod. preukaz typu B
- býva s 3 ďalšími ľudmi
- ...

Jozef

- 53 rokov
- slobodný
- z Myjavy
- VŠ
- vie nemecky
- mal autonehodu
- vod. preukazy typu B,C
- stará sa o rodičov
- ...

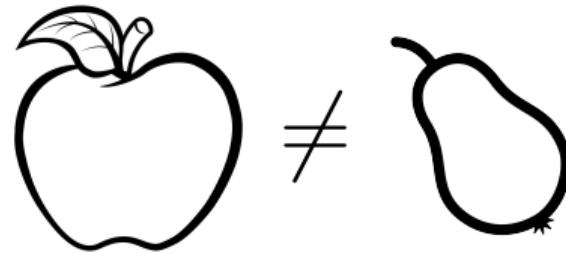


Prekvapivo. Michal a Jozef sú stále rôzni.

Michal \neq Jozef

Prekvapivo. Michal a Jozef sú stále rôzni.

Michal \neq Jozef



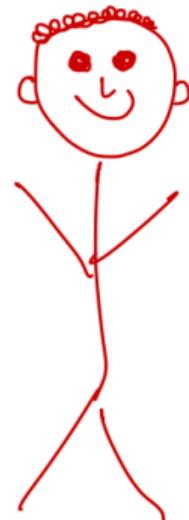
Michal



- 27 rokov
- ženatý
- z Brezna
- stredná škola
- anglicky
- zdravý
- vod. preukaz typu B
- býva v dome s 3 ďalšími ľuďmi
- ...

Peter

- 29 rokov
- slobodný
- z Podbrezovej
- stredná škola
- anglicky, francúzsky
- zdravý
- vod. preukaz typu B
- býva s 2 ľuďmi a so psom
- ...



Prekvapivo. Aj Michal a Peter sú rôzni.

Ale o dosť podobnejší.

Michal \approx Peter

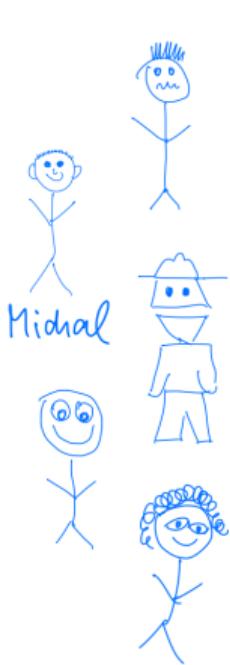
Prekvapivo. Aj Michal a Peter sú rôzni.

Ale o dosť podobnejší.

Michal \approx Peter

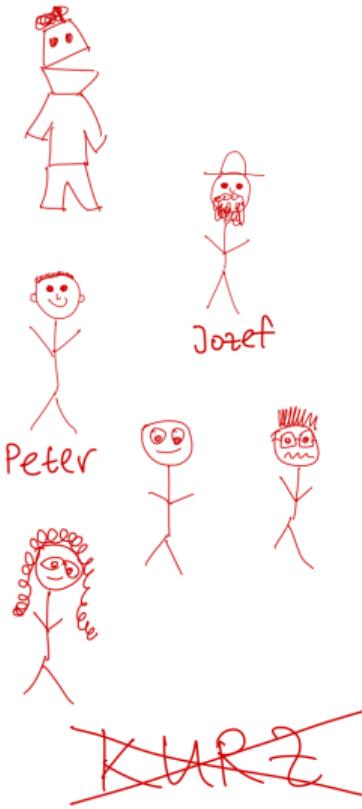


Párovanie (matching)



Mihael

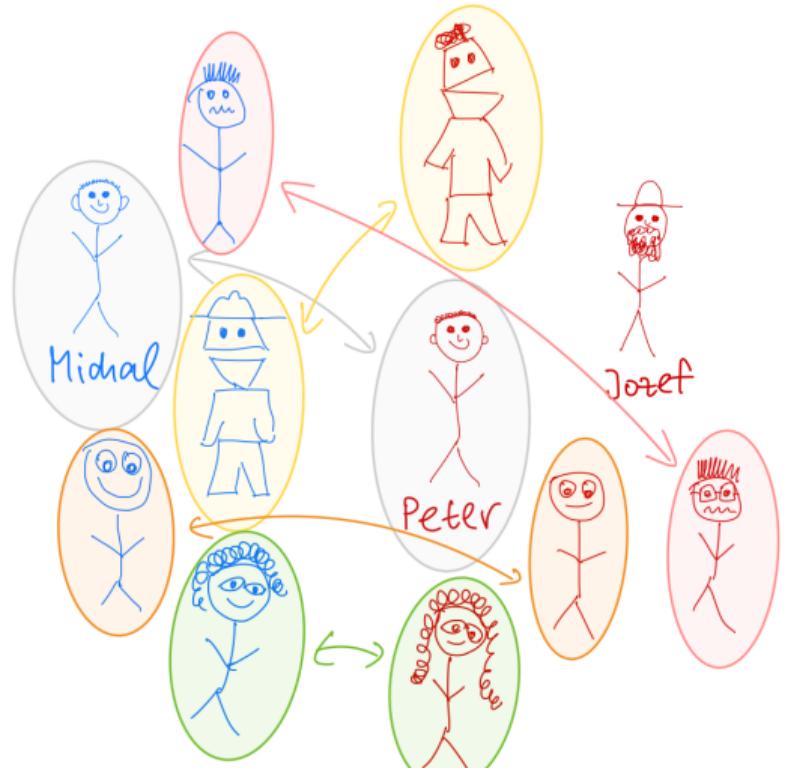
KURZ



Jozef

Peter

~~KURZ~~

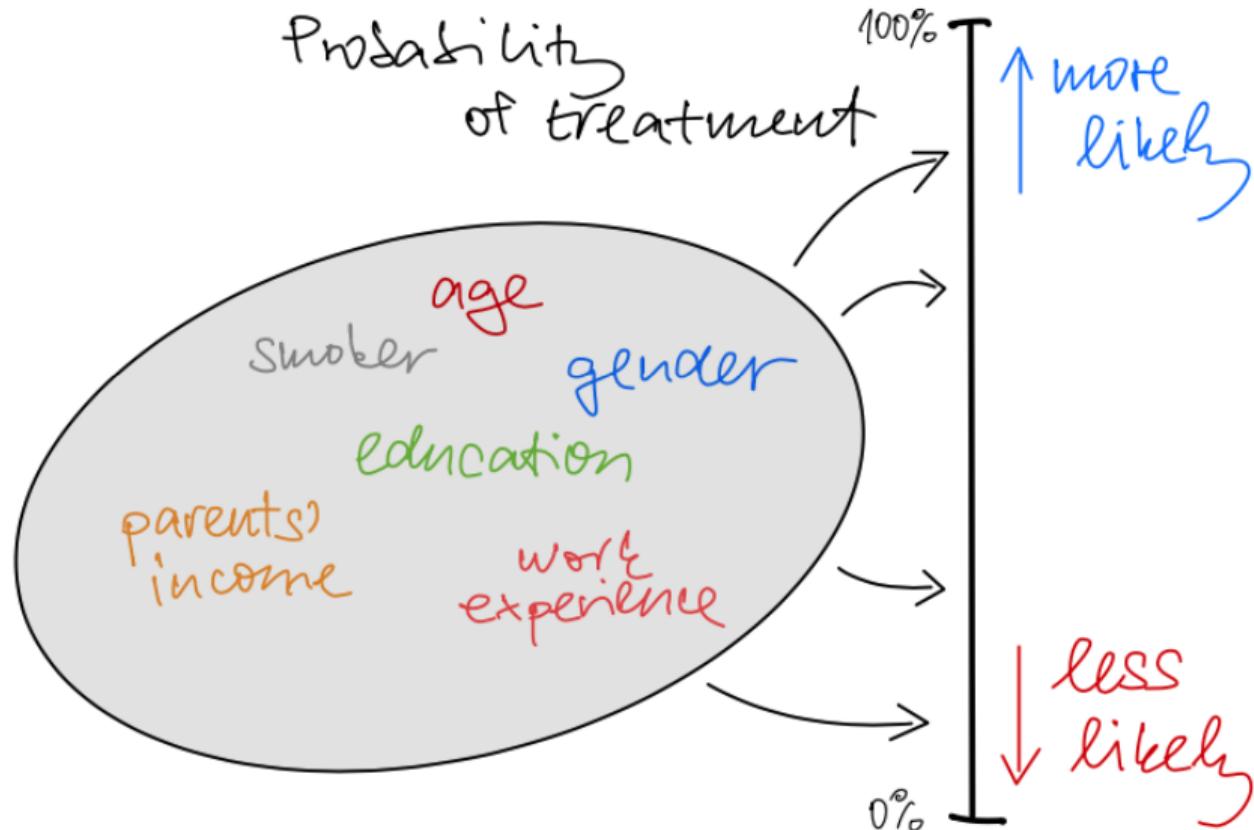


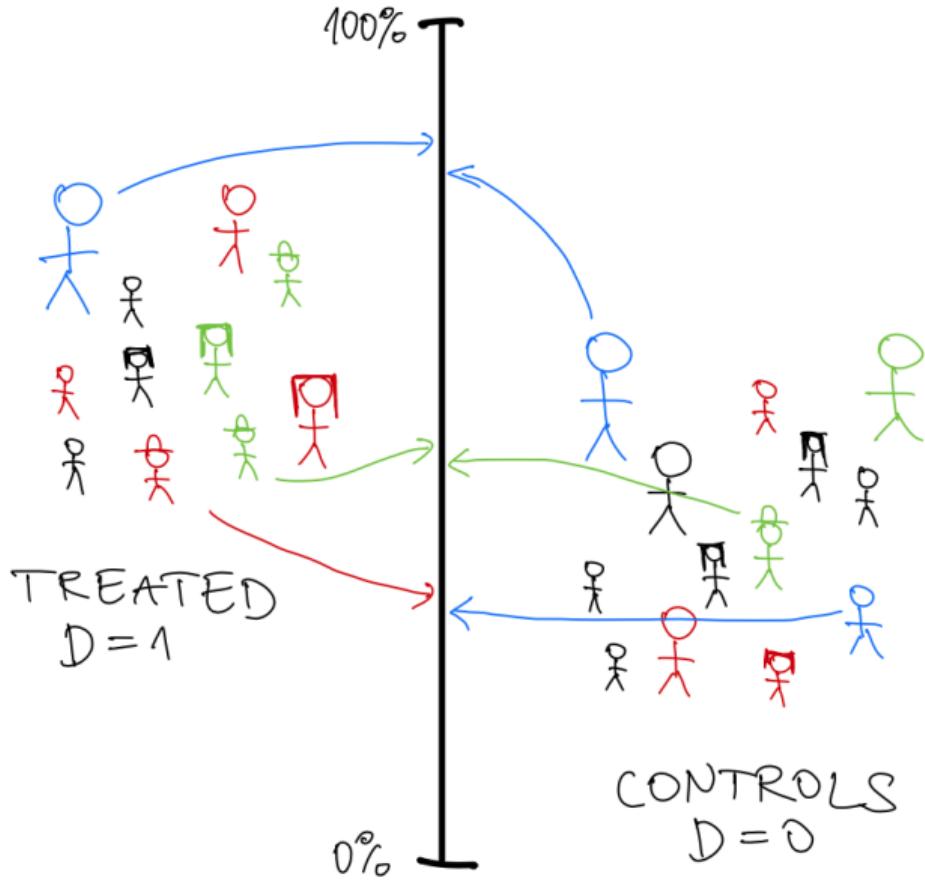
KURT

~~KURT~~

Takýchto **Petrov** je však málo.

A nie každý má svoju "dvojičku" ako **Michal**.





O Michalovi vieme **veľa** informácií.

Ale absolventov kurzu je **málo**.

- vek
 - slobodný/á
 - počet detí
 - mesto
 - typ vzdelania
 - oblasť vzdelania
 - znalosť cudzích jazykov
 - zdravotné znevýhodnenie
 - vod. preukaz
 - história zamestnania
 - typ predošej práce
 - dĺžka predošlého zamestnania
 - klasifikácia predošlého zamestnania
 - počet členov v domácnosti
-
- bariéry zamestnanosti
 - národnosť
 - sociálne dávky
 - zdravotné poistenie
 - zdravotné znevýhodnenie
 - osamelý občan
 - ochota vzdelávať sa
 - ochota dochádzať za prácou
 - účasť na predošom kurze
 - záujem o prácu o neúplný úväzok
 - záujem o prácu v zahraničí
 - úroveň segregácie
 - vzdialenosť od krajského mesta
 - vzdialenosť od Bratislavы
 - ...

Tradičné statistické metódy nevedia pracovať s mnohorozmernými dátami.

Čo s tým?

Priemerný efekt absolvovania kurzu na mzdu

$$\Delta = E[MZDA(\text{kurz}) - MZDA(\cancel{\text{kurz}})]$$

μ

Predikovať priamo

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

μ
→

mzda

μ

Predikovať priamo

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

 $\mu \rightarrow$

mzda

 π

Preváhovanie

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

 $\pi \rightarrow$

ide na
kurz

μ

Predikovať priamo

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

 $\mu \rightarrow$

mzda

 π

Preváhovanie

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

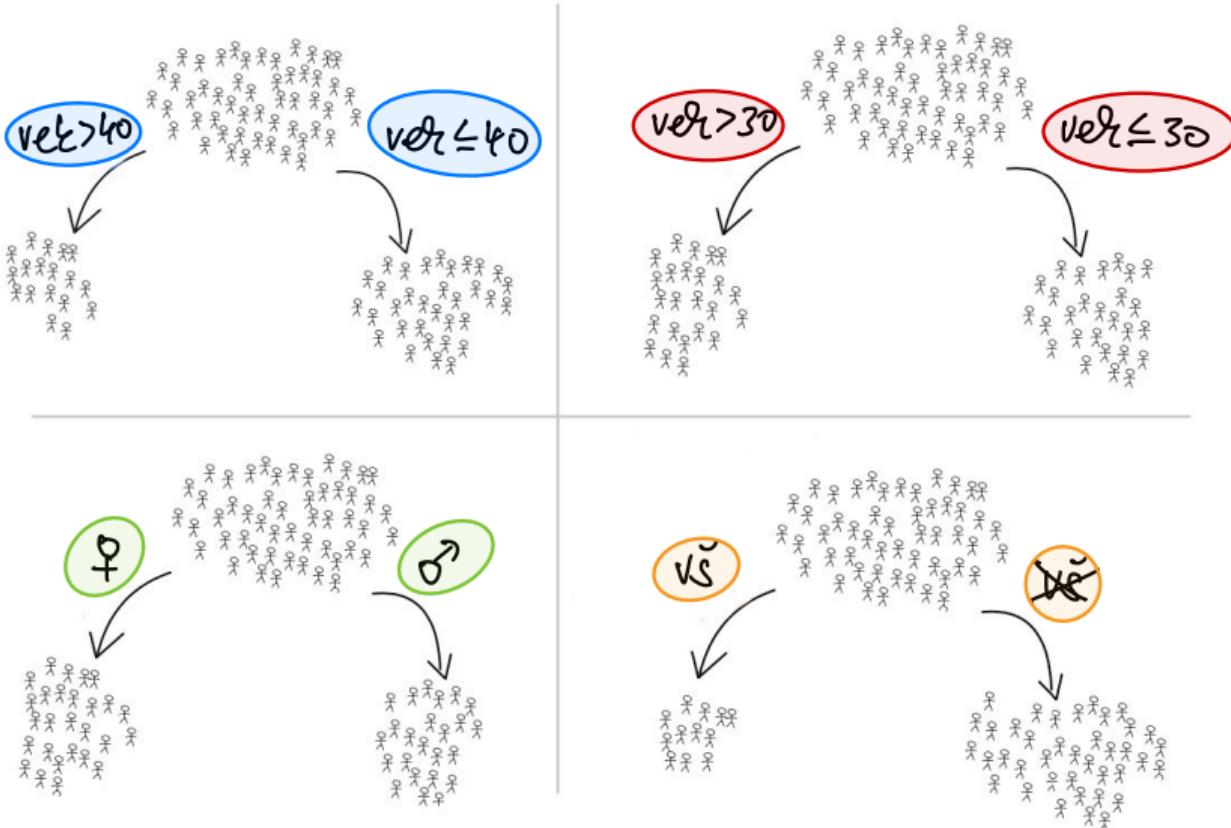
 $\pi \rightarrow$

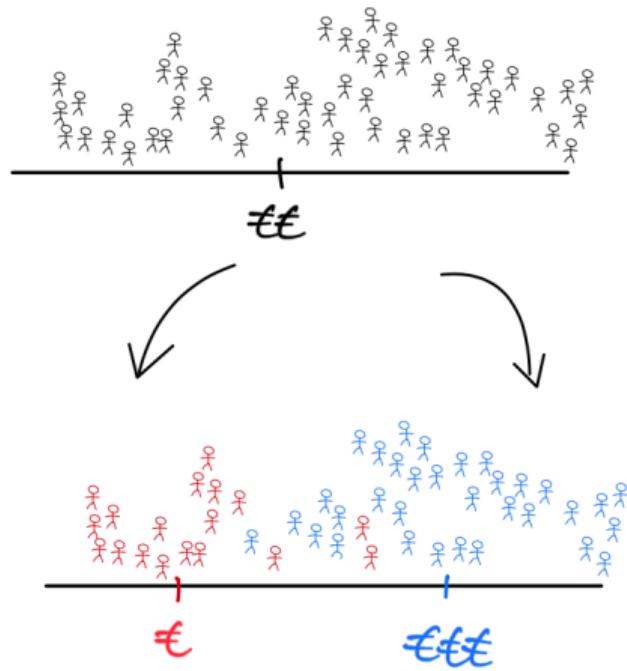
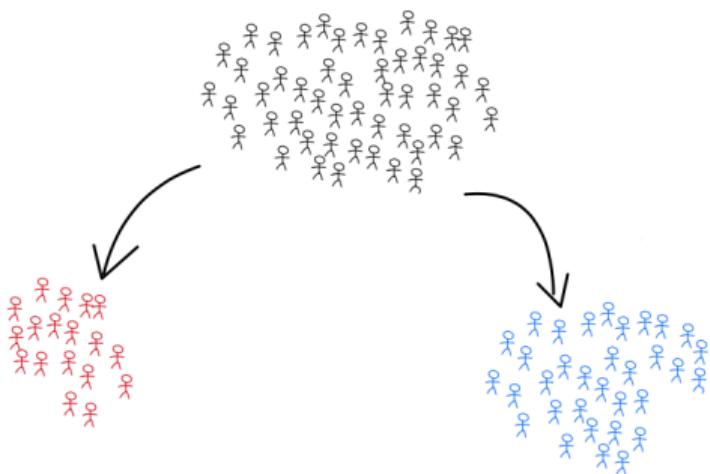
ide na
kurz

MACHINE LEARNING

Aj μ aj π odhadneme pomocou metód strojového učenia.

Príklad algoritmu strojového učenia



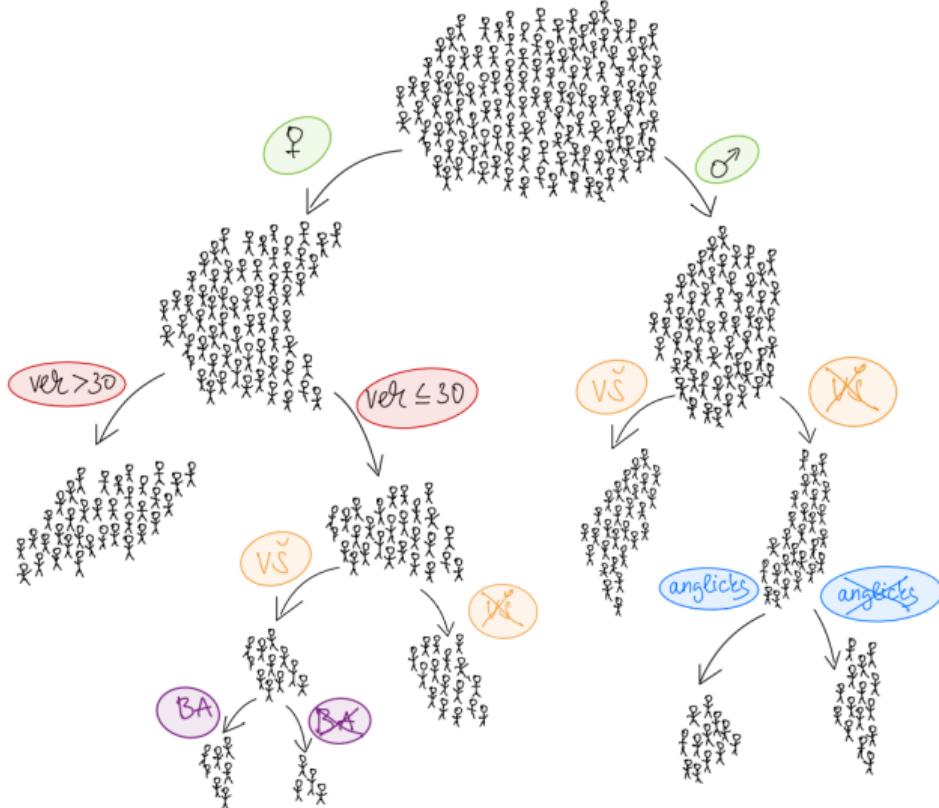


Ako narásť strom?

Dobrá áno/nie otázka?

Vieme to odmerať?

Ako veľký strom?



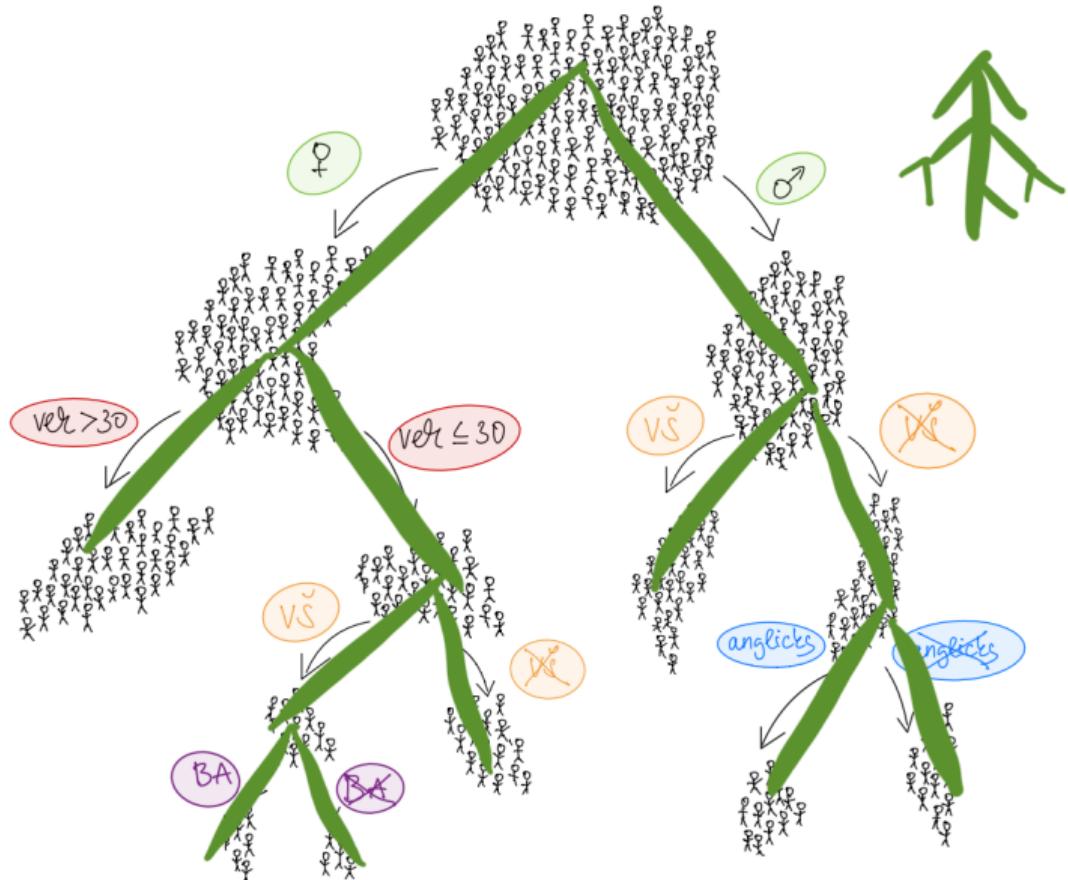
- Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). Classification and Regression Trees.

Ako narásť **strom**?

Dobrá **áno/nie** otázka?

Vieme to odmerať?

Ako veľký strom?



Náhodný les

Pozdravujem vás, lesy, hory,
z tej duše pozdravujem vás!

P. O. Hviezdoslav



P.O.H. v lese.

(Nie až tak) náhodný les

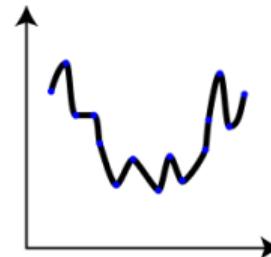
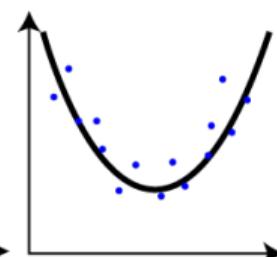
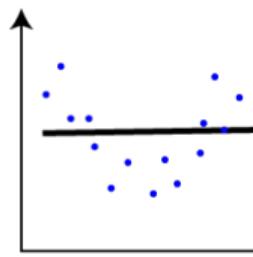
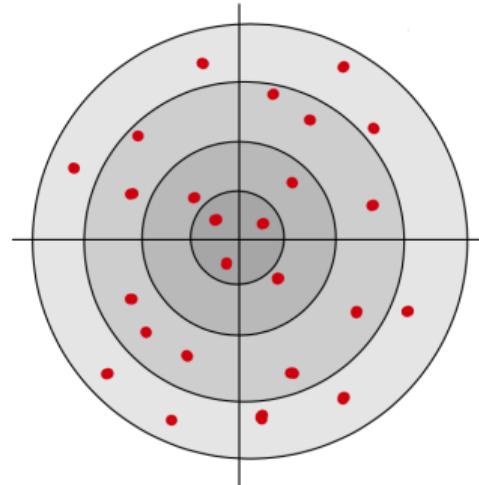
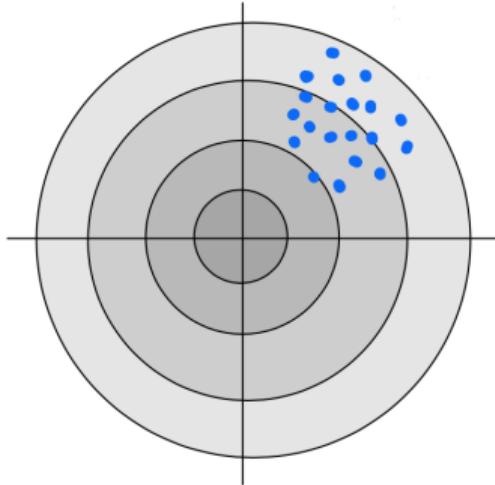


(Už celkom) náhodný les



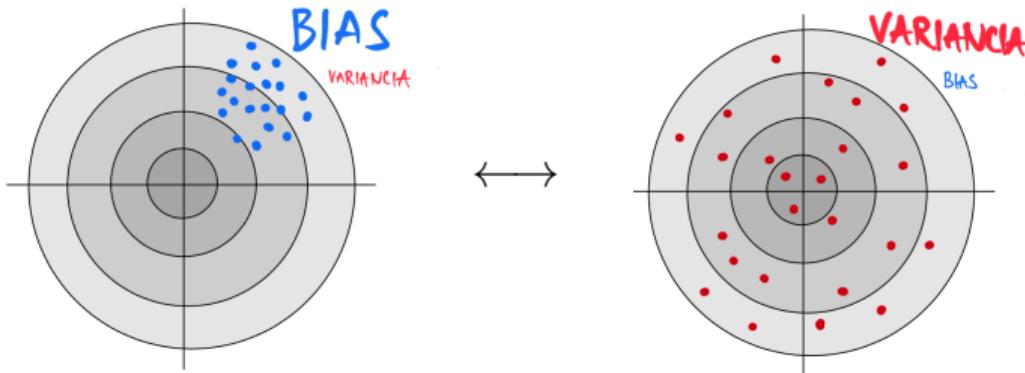
- Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Nevýhody

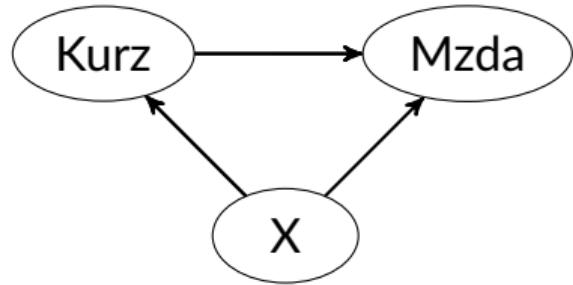


PREDIKČNÁ CHYBA = NÁHODNÁ CHYBA + **NENÁHODNÁ CHYBA**

$$\begin{aligned}\text{PREDIKČNÁ CHYBA} &= \text{NÁHODNÁ CHYBA} + \text{NENÁHODNÁ CHYBA} \\ &= \text{NÁHODNÁ CHYBA} + \underbrace{\text{BIAS}^2 + \text{VARIANCIA}}_{\text{NENÁHODNÁ CHYBA}}\end{aligned}$$



Naspäť k príkladu



X - informácie o veku, pohlaví, vzdelaní, zručnostiach...

$$\Delta = E \left[MZDA(kurz) - MZDA(\cancel{kurz}) \right]$$

- Pearl, J. (2009). Causality. Cambridge university press.
- Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge university press.

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu \longrightarrow$$

$$\Delta\mu$$

efekt na mzdu

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu \longrightarrow$$

$$\Delta_\mu$$

efekt na mzdu

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\pi \longrightarrow$$

$$\Delta_\pi$$

efekt na mzdu

Priama aplikácia μ alebo π (prevážením) na odhad Δ povedie k BIASu.
Obe Δ_μ aj Δ_π konvergujú pomaly: $\Delta_\mu \dashrightarrow \Delta$ a $\Delta_\pi \dashrightarrow \Delta$

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu, \pi \longrightarrow$$

$$\Delta_{\mu, \pi}$$

efekt na mzdu

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu, \pi$$


$$\Delta_{\mu, \pi}$$

efekt na mzdu

$$\Delta_{\mu, \pi} \rightarrow \Delta$$

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, volume 21, pp. C1-C68.
- Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122-129.
- Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu, \pi \longrightarrow$$

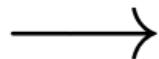
$$\Delta_{\mu, \pi}$$

efekt na mzdu

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

$$\mu, \pi$$



$$\Delta_{\mu, \pi}$$

efekt na mzdu

$$\Delta_{\mu, \pi} \rightarrow \Delta$$

Dve dôležité myšlienky

Model je definovaný pomocou:

$$E[\psi(\text{dáta}, \underbrace{\mu, \pi}_{\equiv \eta}, \Delta)] = 0$$

(1) Neyman orthogonality

- Funkcia ψ je lokálne necitlivá na zmeny v $\eta = (\mu, \pi)$

(2) Sample splitting

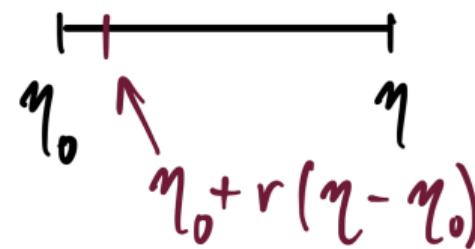
- Rozdelíme dáta a rôzne časti použijeme na odhad $\eta = (\mu, \pi)$ a na odhad Δ

(1) Lokálna necitlivosť (Neyman orthogonality)

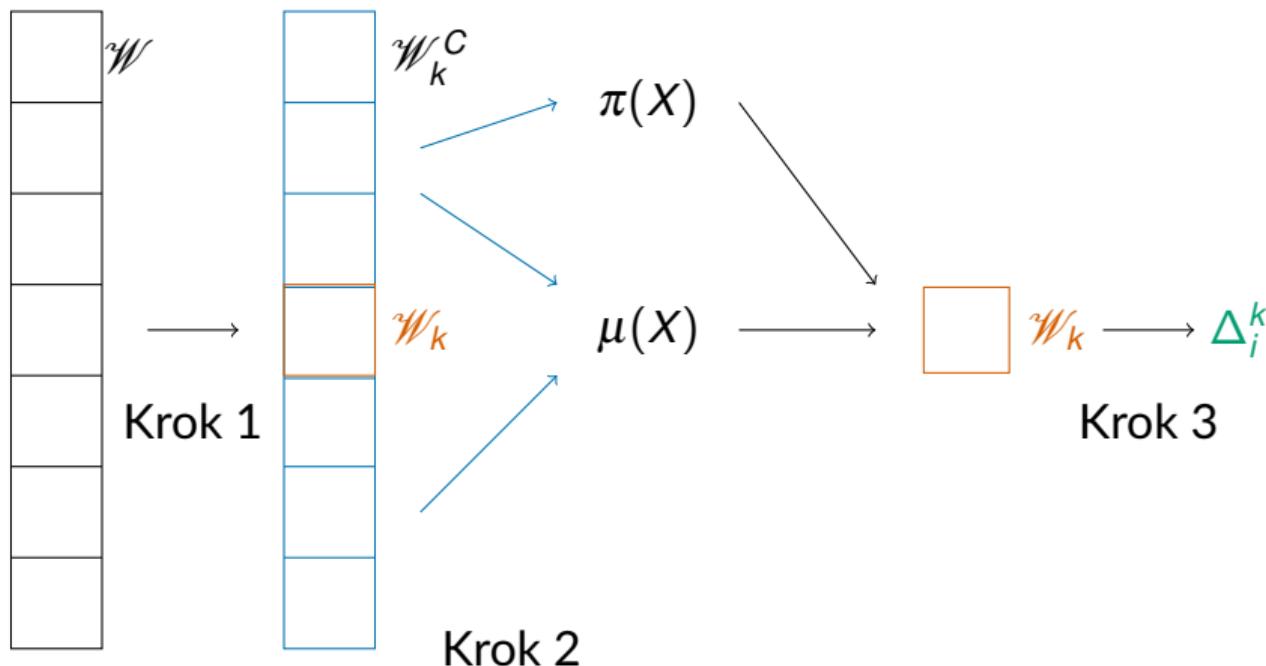
$$E[\psi(\text{dáta}, \underbrace{\mu, \pi, \Delta}_{\equiv \eta})] = 0$$

V blízkom okolí η_0 sa ψ príliš nemení v η :

$$\frac{\partial}{\partial r} E[\psi(\text{dáta}, \eta_0 + r(\eta - \eta_0), \Delta)] \Big|_{r=0} = 0$$



(2) Vyhnut sa overfittingu (Sample splitting)



$$\Delta = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} \Delta_i^k$$

$$\begin{aligned}
 \sqrt{n}(\hat{\Delta} - \Delta) &= \underbrace{a^*}_{\text{Approx. Gaussovský}} + \underbrace{b^*}_{\text{Regularizačný bias}} + \underbrace{c^*}_{\text{Overfitting bias}} \\
 &= \underbrace{a^*}_{\text{Approx. Gaussovský}} + \underbrace{b^*}_{\text{Neyman-orthogonal}} + \underbrace{c^*}_{\text{Sample splitting}} \\
 &\sim N(0, \sigma^2)
 \end{aligned}$$

Príklad

Uvažujme nasledujúci čiastočne lineárny model.

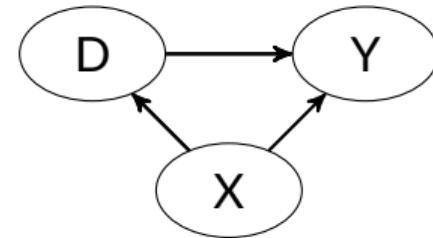
θ je parameter, o ktorý nám ide.

$g(X)$ a $m(X)$ sú flexibilné funkcie, o ktoré nemáme záujem

$$Y = \theta D + g(X) + U, \quad E[U | D, X] = 0,$$
$$D = m(X) + V, \quad E[V | X] = 0.$$

$$\psi = V \cdot U = (D - m(X)) \cdot (Y - g(X) - \theta D)$$

$$E[\psi(\underbrace{Y, X, D}_{\equiv W \sim \text{dáta}}, \underbrace{m, g, \theta}_{\equiv \eta})] = 0$$



Detailedly

Double machine learning - zhrnutie

- Estimátor $\hat{\Delta}$ založený na Neyman-ortogonálnej funkcií ψ
- Použijúc sample splitting
- Estimátory μ and π sú "dostatočne dobré" (e.g. konvergujú rýchlosťou aspoň $n^{-1/4}$)

Double machine learning - zhrnutie

- Estimátor $\hat{\Delta}$ založený na Neyman-ortogonálnej funkcií ψ
- Použijúc sample splitting
- Estimátory μ and π sú "dostatočne dobré" (e.g. konvergujú rýchlosťou aspoň $n^{-1/4}$)

Veta 1 (Chernozhukov et al. 2018):

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

Estimátor $\hat{\Delta}$ je asymptoticky normálne rozdelený a je \sqrt{n} -konzistentný.

Double machine learning - zhrnutie

DML nám poskytuje odhadovacie prístupy, ktoré:

Double machine learning - zhrnutie

DML nám poskytuje odhadovacie prístupy, ktoré:

- zvládajú veľadimenzionálne dáta

Double machine learning - zhrnutie

DML nám poskytuje odhadovacie prístupy, ktoré:

- zvládajú veľadimenzionálne dátá
- sú flexibilné

Double machine learning - zhrnutie

DML nám poskytuje odhadovacie prístupy, ktoré:

- zvládajú veľadimenzionálne dátá
- sú flexibilné
- vedia využiť prediktívne vlastnosti ML

Double machine learning - zhrnutie

DML nám poskytuje odhadovacie prístupy, ktoré:

- zvládajú veľadimenzionálne dátá
- sú flexibilné
- vedia využiť prediktívne vlastnosti ML
- majú dobré štatistické vlastnosti

Double/debiased machine learning for treatment and structural parameters

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 4471 Related articles All 30 versions

[PDF] oup.com



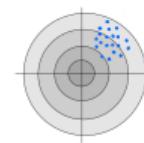
Limitácie - "Kitchen sink" regresia



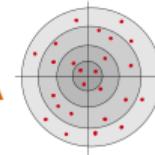
Hünermund, Beyers and Caspi (2023)

Zatial'

- Kauzalita je náročná
- O to viac v modeloch s veľa informáciami



BIAS vs VARIANCIA



Zatial'

- Kauzalita je náročná
- O to viac v modeloch s veľa informáciami



- ML metódy sú skvelé na predikciu
- Na odhadovanie parametrov už nie

$$\Delta_{\mu} \dashrightarrow \Delta \quad \text{a} \quad \Delta_{\pi} \dashrightarrow \Delta$$

Zatial'

- Kauzalita je náročná
- O to viac v modeloch s veľa informáciami



- ML metódy sú skvelé na predikciu
- Na odhadovanie parametrov už nie

$$\Delta_{\mu} \dashrightarrow \Delta \quad \text{a} \quad \Delta_{\pi} \dashrightarrow \Delta$$

- Skombinujeme dve slabé ML metódy μ, π
- Dostaneme **odhad** s dobrými vlastnosťami

$$\Delta_{\mu, \pi} \rightarrow \Delta$$

Rozšírenia

- ...
- DATA
- ...

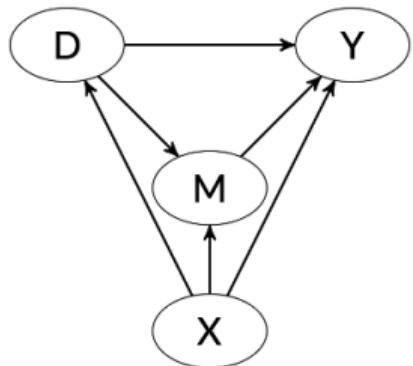
μ, π
→

$\Delta_{\mu, \pi}$
efekt

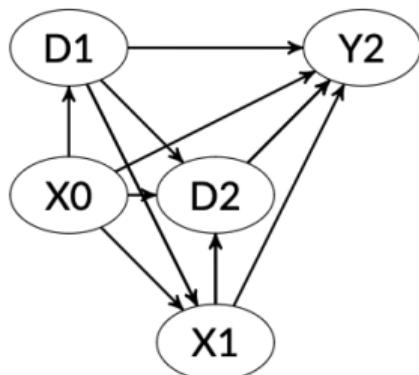
???

$\Delta_{\mu, \pi} \rightarrow \Delta$

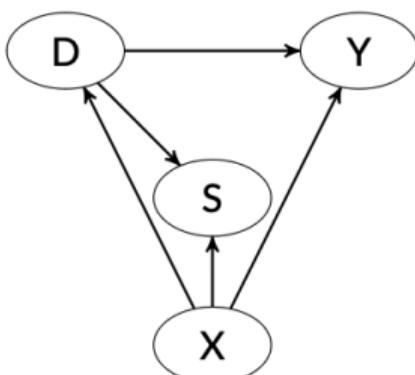
Mediačná analýza



Dynamické efekty



Výberová vzorka



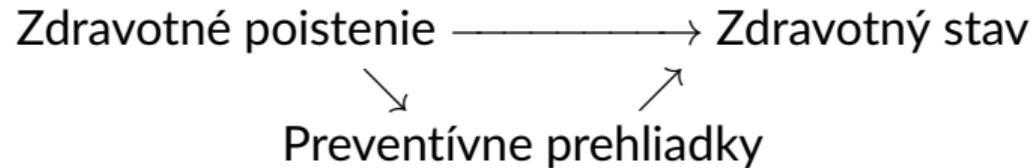
- Mediačná analýza (H. Farbmacher, M. Huber, H. Langen, LL, M. Spindler)
- Dynamické efekty (H. Bodory, M. Huber, LL)
- Výberová výchylka (M. Bia, M. Huber, LL)

Prvé rozšírenie

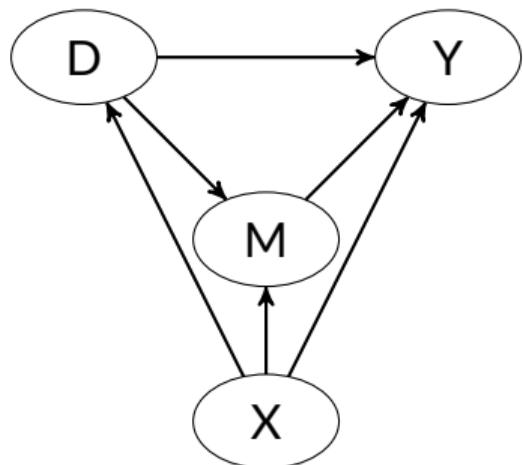
DML and mediation analysis

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)

Príklad



DML a mediačná analýza



Objekt záujmu:

Nepriamy efekt: $E[Y(d, M(1)) - Y(d, M(0))]$

Priamy efekt: $E[Y(1, M(d)) - Y(0, M(d))]$

Identifikačné predpoklady:

1) Podmienená nezávislosť D

$$\{Y(d', m), M(d)\} \perp\!\!\! \perp D \mid X$$

2) Podmienená nezávislosť M

$$Y(d', m) \perp\!\!\! \perp M \mid D = d, X = x$$

3) Spoločný nosič

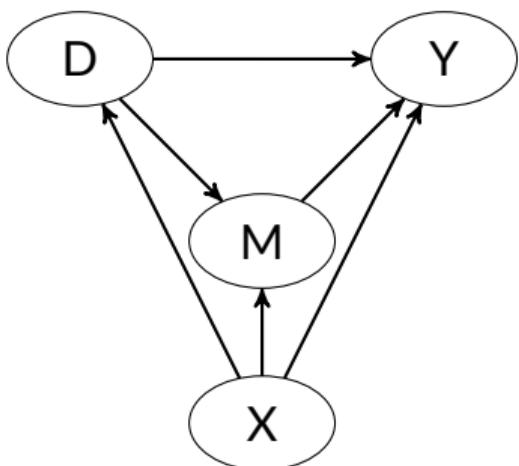
$$\Pr(D = d \mid M = m, X = x) > 0$$

DML a mediačná analýza

Model:

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

$$E[\psi(W; \theta_0, \eta)] = E[Y(d, M(1-d))] - \theta_0 = 0$$



Dáta: $W = (Y, D, M, X)$

Parametre vedľajšieho záujmu: $\eta = (p_d, p_{dm}, \mu, \omega)$

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

Aplikácia

Výsledky:

	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
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

- Zdravotné poistenie mierne zlepšuje všeobecné zdravie v krátkodobom horizonte u mladých dospelých v USA mechanizmami inými než bežné preventívne prehliadky.

Detaily

Druhé rozšírenie

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

Príklad

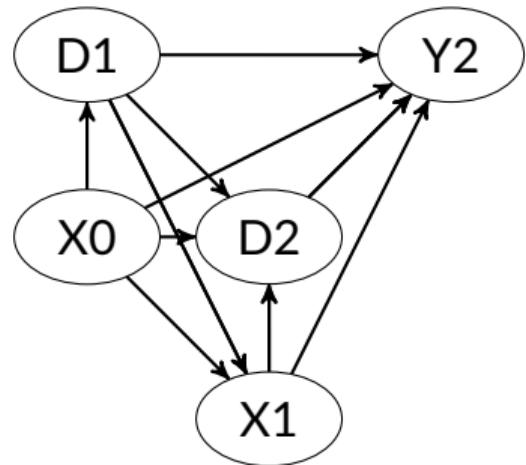
Rekvalifikačné (akademické/odborné) kurzy → Zamestnanosť

Detaily

DML a dynamické efekty

Objekt záujmu:

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



Identifikačné predpoklady:

1) Podmienená nez. prvej intervencie

$$Y_2(\underline{d}_2) \perp D_1 | X_0, \text{ for } \underline{d}_2 \in \{0, 1, \dots, Q\}^2$$

2) Podmienená nez. druhej intervencie

$$Y_2(\underline{d}_2) \perp D_2 | D_1, X_0, X_1, \text{ for } \underline{d}_2 \in \{0, 1, \dots, Q\}^2.$$

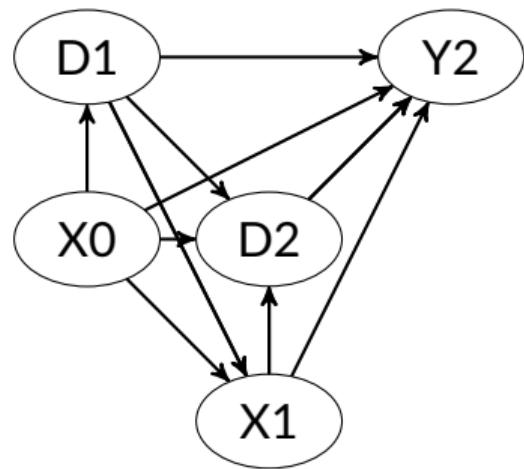
3) Spoločný nosič

$$\Pr(D_1 = d_1 | X_0) > 0, \Pr(D_2 = d_2 | D_1, X_1) > 0$$

DML a dynamické efekty

Model:

$$\begin{aligned}\psi(W; \theta_0, \eta) &= \frac{I\{D_1 = d_1\} \cdot I\{D_2 = d_2\} \cdot [Y_2 - \mu^{Y_2}(d_2, X_1)]}{p^{d_1}(X_0) \cdot p^{d_2}(d_1, X_1)} \\ &+ \frac{I\{D_1 = d_1\} \cdot [\mu^{Y_2}(d_2, X_1) - v^{Y_2}(d_2, X_0)]}{p^{d_1}(X_0)} + v^{Y_2}(d_2, X_0) - \theta_0.\end{aligned}$$



Dáta: $W = (Y_2, D_1, D_2, X_0, X_1)$

Parametre vedľajšieho záujmu:

$$\eta = (p^{d_1}, p^{d_2}, \mu^{Y_2}, v^{Y_2})$$

- $p^{d_1}(X_0) \equiv \Pr(D_1 = d_1 | X_0)$
- $p^{d_2}(D_1, X_1) \equiv \Pr(D_2 = d_2 | D_1, X_1)$
- $\mu^{Y_2}(D_2, X_1) \equiv E[Y_2 | D_2, X_0, X_1]$
- $v^{Y_2}(D_2, X_0) \equiv E[E[Y_2 | D_2, X_0, X_1] | D_1, X_0],$

DML a dynamické efekty: Aplikácia

Výsledky (zamestnanosť po 4 rokoch):

\underline{d}_2	\underline{d}_2^*	$\hat{F}[Y_2(\underline{d}_2^*) S = 1]$	$\hat{\Delta}(\underline{d}_2, \underline{d}_2^*, S = 1)$	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

3 = vocational 2 = academic
1 = no training
10% ↑ employment after 4 years

Detaily

Tretie rozšírenie

DML and sample selection models

Michela Bia, Martin Huber and Lukáš Lafférs: Double machine learning for sample selection models. Journal of Business & Economic Statistics, 2024, 42 (3), 958-969

Príklad

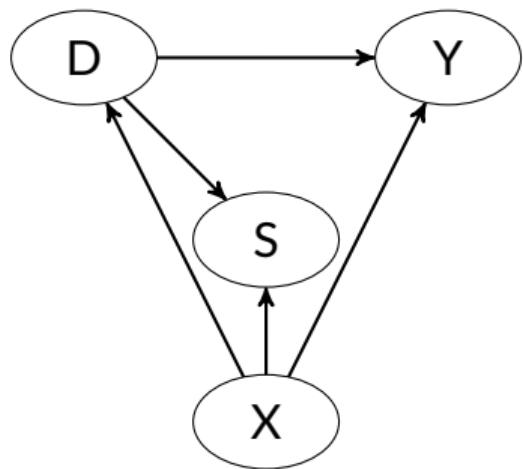
Rekvalifikačné (akademické/odborné) kurzy → Mzdy

Detaily

DML a modely s výberovou výchylkou

Objekt záujmu:

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



Identifikačné predpoklady

1) Podmienená nez. intervencie:

$$Y(d) \perp\!\!\! \perp D | X = x$$

2) Podmienená nez. selekcie

$$Y \perp\!\!\! \perp S | D = d, X = x$$

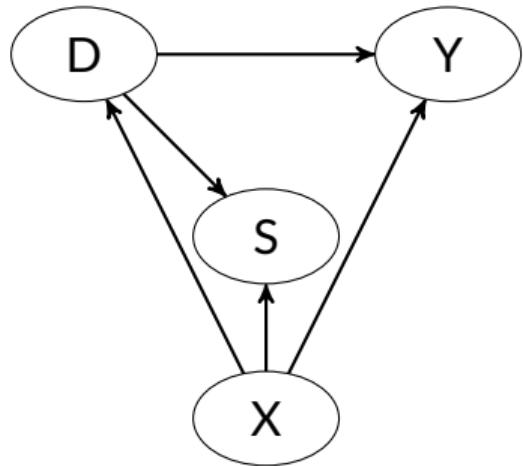
3) Spoločný nosič

(a) $\Pr(D = d | X = x) > 0$ and (b)

$$\Pr(S = 1 | D = d, X = x) > 0$$

DML a modely s výberovou výchylkou

Model:



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

Dáta: $W = (Y, S, S, D, X)$

Parametre vedľajšieho záujmu: $\eta = (p^d, \pi, \mu)$

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

DML a modely s výberovou výchylkou: Aplikácia

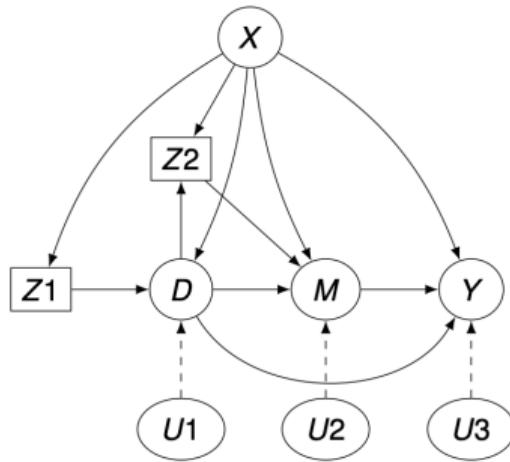
$D = 1$	$D = 0$	ATE	štandardná odchýlka	p-hodnota
Veta 1 (MAR)				
akademické	bez kurzu	-0.683	1.073	0.524
odborné	bez kurzu	0.611	0.629	0.331
Veta 3 (IV)				
akademické	bez kurzu	-0.631	1.052	0.549
odborné	bez kurzu	0.586	0.645	0.364
Veta 4 (sekvenčné)				
akademické	bez kurzu	0.149	0.199	0.454
odborné	bez kurzu	0.567	0.208	0.007

Pozorujeme mierne zvýšenie dlhodobejších hodinových miezd.

Detaily

Iné

Testovateľné predpoklady (s M. Huberom, K. Kloiberom)



$$Y \perp\!\!\!\perp Z_1 | D = d, X = x, \quad (T\text{Ia})$$

$$M \perp\!\!\!\perp Z_1 | D = d, X = x, \quad (T\text{Ib})$$

$$Y \perp\!\!\!\perp Z_2 | D = d, M = m, X = x \quad (T\text{Ic})$$

Under $\underbrace{1, 4, 5}_{\text{causal structure}} + \text{relevance conditions}$

$\underbrace{6a, 6b, 7}_{\text{Sequential ignorability}}, \underbrace{8a, 8b, 9}_{\text{Instruments}}$

$\iff \underbrace{(T\text{Ia}), (T\text{Ib}), (T\text{Ic})}_{\text{Testable implications}}$

- Citlivosť na výber hyperparametrov ML (Ahrens, A., Hansen, C.B., Schaffer, M.E. and Wiemann, 2025)
- Explicitné adresovania heterogeneity efektov (Athey and Wager (2018))
- Férivost dostupnosti programu vzhľadom na citlivé informácie (Bearth, N., Lechner, M., Mareckova, J., and Muny, F. (2025))
- Prehľadávanie kauzálnych štruktúr (grafov) kompatibilných s dátami (Causal discovery)

Rekapitulácia

Rekapitulácia

DML je užitočný rámec pre odhadovanie v prostredí s vysokou dimenzionalitou.

Rekapitulácia

DML je užitočný rámec pre odhadovanie v prostredí s vysokou dimenzionalitou.

Dokáže automaticky vyberať medzi mnohými premennými a vyhnúť sa **regularizačnému biasu** (cez Neymanovu ortogonalitu) a **overfitting biasu** (cez cross-fitting) a

poskytnúť odhad, ktorý je root-n konzistentný a asymptoticky normálny.

Rekapitulácia

DML je užitočný rámec pre odhadovanie v prostredí s vysokou dimenzionalitou.

Dokáže automaticky vyberať medzi mnohými premennými a vyhnúť sa **regularizačnému biasu** (cez Neymanovu ortogonalitu) a **overfitting biasu** (cez cross-fitting) a

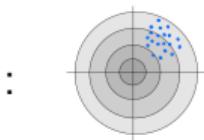
poskytnúť odhad, ktorý je root-n konzistentný a asymptoticky normálny.

Ukázal som tri rozšírenia DML, ktoré sa javia ako empiricky relevantné a užitočné.

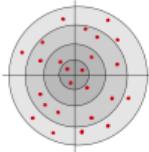
Implementované v balíku causalweight pre R (Bodory a Huber 2018)

 \neq  $\xrightarrow{\text{Párovanie}}$  \approx 

 \neq  $\xrightarrow{\text{Párovanie}}$  \approx 



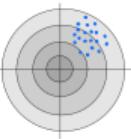
BIAS vs VARIANCIA



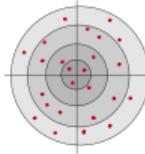
 \neq  $\xrightarrow{\text{Párovanie}}$  \approx 



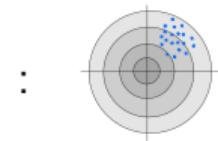
Δ_{μ} \dashrightarrow 
 Δ_{π} \dashrightarrow 



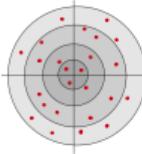
BIAS vs VARIANCIA



$$\text{apple} \neq \text{pear} \xrightarrow{\text{Párovanie}} \text{apple} \approx \text{apple}$$



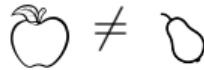
BIAS vs VARIANCIA



$$\Delta_{\mu} \dashrightarrow \Delta$$
$$\Delta_{\pi} \dashrightarrow \Delta$$

Double ML

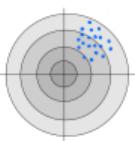
$$\Delta_{\mu, \pi} \rightarrow \Delta$$



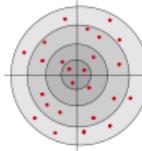
Párovanie



:



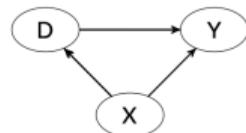
BIAS vs VARIANCIA



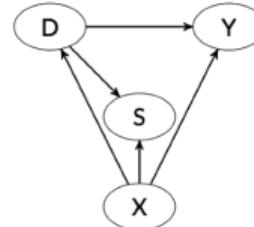
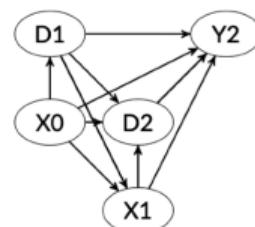
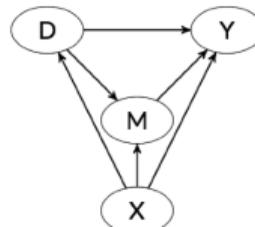
$$\begin{array}{c} \Delta_\mu \dashrightarrow \Delta \\ \Delta_\pi \dashrightarrow \Delta \end{array}$$

Double ML

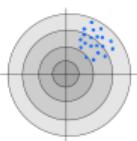
$$\Delta_{\mu, \pi} \rightarrow \Delta$$



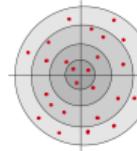
Rozšírenia



 \neq  $\xrightarrow{\text{Párovanie}}$  \approx 



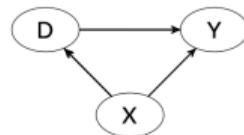
BIAS vs VARIANCIA



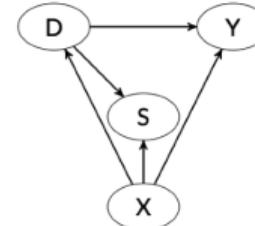
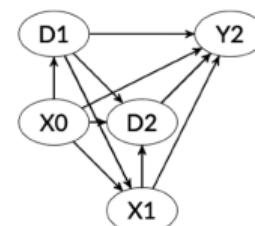
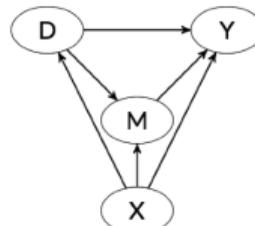
$$\begin{array}{c} \Delta_\mu \dashrightarrow \Delta \\ \Delta_\pi \dashrightarrow \Delta \end{array}$$

$\xrightarrow{\text{Double ML}}$

$$\Delta_{\mu, \pi} \rightarrow \Delta$$



$\xrightarrow{\text{Rozšírenia}}$



Ďakujem za pozornosť.

lukaslaffers.github.io

#UMBmath

References

- Inštrumentálne premenné Imbens, Guido W., and Joshua D. Angrist. (1994). *Identification and Estimation of Local Average Treatment Effects*. *Econometrica*, 62(2), 467–475.
- Dizajn regresnej diskontinuity Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. (2001). *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design*. *Econometrica*, 69(1), 201–209.
- Metódā syntetickej kontrolnej premennej Abadie, Alberto, and Javier Gardeazabal. (2003). *The Economic Costs of Conflict: A Case Study of the Basque Country*. *American Economic Review*, 93(1), 113–132.
- Metódā rozdielu v rozdieloch (difference-in-differences) Card, David, and Alan B. Krueger. (1994). *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*. *American Economic Review*, 84(4), 772–793.
- 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).
- Ahrens, A., Hansen, C. B., Schaffer, M. E., and Wiemann, M. (2025). *On the Sensitivity of Machine Learning Methods to Hyperparameter Choice*. Journal/Publisher TBD.
- Athey, S., and Wager, S. (2018). *Estimating Treatment Effects with Causal Forests*. *Journal of the American Statistical Association*, 113(523), 1228–1242.
- Bearth, N., Lechner, M., Mareckova, J., and Muny, F. (2025). Fairness-Aware and Interpretable Policy Learning. arXiv preprint arXiv:2509.12119.
- Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.

Neyman-ortogonalita pre ψ

Overíme, že ψ spĺňa podmienku Neymanovej ortogonality.

Označenie:

- $\eta = (m, g)$ je vektor vedľajších (nuisance) parametrov, $\eta_0 = (m_0, g_0)$ je skutočný vektor,
- $\eta_r = \eta_0 + r(\eta - \eta_0)$.

Neyman-ortogonalita pre ψ

$$\begin{aligned}\psi(W; \eta_r, \theta_0) &= (D - m_0(X) - r(m(X) - m_0(X))) \cdot (Y - g_0(X) - r(g(x) - g_0(X)) - D\theta_0) \\ &= (D - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) + \\ &\quad - r(D - m_0(X)) \cdot (g(x) - g_0(X)) \\ &\quad - r(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) \\ &\quad + r^2(m(X) - m_0(X)) \cdot (g(x) - g_0(X))\end{aligned}$$

Neyman-ortogonalita pre ψ

$$\begin{aligned}\psi(W; \eta_r, \theta_0) &= (D - m_0(X) - r(m(X) - m_0(X))) \cdot (Y - g_0(X) - r(g(x) - g_0(X)) - D\theta_0) \\ &= (D - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) + \\ &\quad - r(D - m_0(X)) \cdot (g(x) - g_0(X)) \\ &\quad - r(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) \\ &\quad + r^2(m(X) - m_0(X)) \cdot (g(x) - g_0(X))\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial r} E[\psi(W; \eta_r, \theta_0)] &= -E[(D - m_0(X)) \cdot (g(x) - g_0(X))] \\ &\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)] \\ &\quad + 2 \cdot r \cdot E[(m(X) - m_0(X)) \cdot (g(x) - g_0(X))]\end{aligned}$$

Neyman-ortogonalita pre ψ

$$\begin{aligned}\psi(W; \eta_r, \theta_0) &= (D - m_0(X) - r(m(X) - m_0(X))) \cdot (Y - g_0(X) - r(g(x) - g_0(X)) - D\theta_0) \\ &= (D - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) + \\ &\quad - r(D - m_0(X)) \cdot (g(x) - g_0(X)) \\ &\quad - r(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) \\ &\quad + r^2(m(X) - m_0(X)) \cdot (g(x) - g_0(X))\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial r} E[\psi(W; \eta_r, \theta_0)] &= -E[(D - m_0(X)) \cdot (g(x) - g_0(X))] \\ &\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)] \\ &\quad + 2 \cdot r \cdot E[(m(X) - m_0(X)) \cdot (g(x) - g_0(X))]\end{aligned}$$

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

Neyman-ortogonalita pre ψ

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

Neyman-ortogonalita pre ψ

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

nakoľko

Neyman-ortogonalita pre ψ

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

nakolko

$$\mathbb{E}[(D - m_0(X)) \cdot (g(X) - g_0(X))] = \mathbb{E}[(g(X) - g_0(X)) \cdot \underbrace{\mathbb{E}[D - m_0(X)|X]}_{\mathbb{E}[V|X]=0}] = 0$$

Neyman-ortogonalita pre ψ

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

nakonko

$$\mathbb{E}[(D - m_0(X)) \cdot (g(X) - g_0(X))] = \mathbb{E}[(g(X) - g_0(X)) \cdot \underbrace{\mathbb{E}[D - m_0(X)|X]}_{\mathbb{E}[V|X]=0}] = 0$$

$$\mathbb{E}[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)] = \mathbb{E}[(m(X) - m_0(X)) \cdot \underbrace{\mathbb{E}[Y - g_0(X) - D\theta_0|X]}_{\mathbb{E}[U|X,D]=0}] = 0$$

a teda ψ je vskutku Neyman-ortogonálna.

Späť

Súvisiace s hodnotením pracovného tréningu:



- Priamy efekt na mzdy programu **Job Corps** s využitím pracovných skúseností ako mediátora (Flores a Flores-Lagunes (2009))
- Efekt **Perry Preschool Program** na zdravé správanie sprostredkovaný osobnostnými črtami (Conti, Heckman a Pinto (2016))
- Aký je efekt dôkladnejších poradcov v procese poradenstva na zamestnanosť sprostredkovaný zaradením do programu trhu práce (Huber, Lechner a Mellace (2017))

Súvisiace so vzdelaním a mzdami:



- Ako **rast v chudobe** ovplyvňuje **ekonomické výsledky** v dospelosti s využitím **vzdelania** ako mediátora (Bellani a Bia (2018))
- Decompozícia mzdového rozdielu (**pohlavie**, **sociálno-ekonomicke premenné**, **mzda**) (Huber (2015))
- Efekt **vzdelania** na **úmrtnosť** sprostredkovany **kognitívnymi schopnosťami** (Bijwaard a Jones (2018))

Na základe nástrojových premenných:



- Efekt **vzdelania** na **spokojnosť so životom** s využitím **príjmu** ako mediátora (Powdthavee, Lekfuangfu a Wooden (2013))
- Efekt **vzdelania** na **zdravie** sprostredkovaný **zdravotným správaním** (Brunello, Fort, Schneeweis a Winter-Ebmer (2016))
- Efekt **zloženia rodiny** na **vzdelanie prvého dieťaťa** s využitím **veľkosti rodiny** ako mediátora (Chen, Chen a Liu (2017))

Späť

Mediáčná analýza príklad - detailey

- zdravotné poistenie → zdravotný stav
- zdravotné poistenie → pravidelné preventívne prehliadky → zdravotný stav
- X - demografia, rodinné zázemie, vzdelanie, trh práce, charakteristiky domácnosti, duševné zdravie, výživa, fyzická aktivita....
(755 kontrolných premenných, z roku 2005)
- Národný longitudinálny prieskum mládeže 1997 (NLSY97), prieskum amerického Ministerstva práce (2019) ($n \approx 7500$)
- väčšina štúdií zistila významný efekt na konkrétny typ skríningu (rakovina, mŕtvica...)
- máme mladších respondentov a krátkodobé efekty (2006 → 2007 → 2008)
- zdravie - „výborné“ až „zlé“, negatívny ATE \approx zlepšenie

Dynamický príklad - detaile

- Tréning → Zamestnanie (po 4 rokoch)
- X - 1184 premenných (X_0 – 814 , X_1 – 374) socio-ekonomickej charakteristiky, vzdelanie a tréning pred zásahom, história trhu práce, aktivity hľadania zamestnania, prijímanie sociálnych dávok, zdravie, kriminalita...
- Job Corps poskytuje odborné školenie a akademické vyučovanie pre znevýhodnené osoby vo veku 16 až 24 rokov
- V súčasnosti približne 50 000 účastníkov každý rok.
- Vzorka pochádza z experimentálnej štúdie Job Corps realizovanej v polovici 90. rokov, Schochet et al. (2008): 11313 mladých osôb, s ukončenými rozhovormi štyri roky po randomizácii (6828 zaradených do Job Corps, 4485 randomizovaných mimo).
- Sekvencie zásahov sú založené na účasti v akademickom alebo odbornom tréningu v prvom alebo druhom roku po randomizácii medzi tými, ktorí boli randomizovaní do programu.

Späť

Príklad výberu vzorky - detaile

- Tréning → Hodinová mzda
- Stovky počiatočných kovariátov X (socio-ekonomicke premenné, história trhu práce, kriminalita, zdravie...).
- Job Corps poskytuje odborné školenie a akademické vyučovanie pre znevýhodnené osoby vo veku 16 až 24 rokov
- V súčasnosti približne 50 000 účastníkov každý rok.
- Vzorka pochádza z experimentálnej štúdie Job Corps ($n \approx 3600$)
- Výsledok Y je **hodinová mzda** v poslednom týždni prvého roka alebo štyri roky po randomizácii, pozorovaný pod podmienkou zamestnania S .
- Zásah D je účasť v akademickom alebo odbornom **školení** v prvom roku po randomizácii medzi tými, ktorí boli randomizovaní do programu.

Späť