

Balanced Group Average Treatment Effects

Verein für Socialpolitik: Jahrestagung 2023

Nora Beath & Michael Lechner

September 25, 2023

Motivation and Related Literature

- **Detecting heterogeneity:**

- One of the great promises of causal machine learning literature
- Many applied papers use methods to detect heterogeneities (e.g. Cockx, Lechner, & Bollens, 2023; Davis & Heller, 2017; Knaus, Lechner, & Strittmatter, 2022)
 - e.g. heterogeneous effects by gender, age, education

- **Interpretation of detected heterogeneity:**

- Not clear how to interpret effect heterogeneity if we compare two groups with each other
- Is effect heterogeneity only correlated with a particular variable or really caused by it?

Heterogeneous treatment effects:

- Causal forests
 - e.g. Athey, Tibshirani, & Wager (2019); Lechner & Mareckova (2022)
- Using doubly-robust score functions
 - with a linear second step:
e.g. Semenova & Chernozhukov (2021)
 - with a nonparametric second step:
e.g. Kennedy (2020); Zimmert & Lechner (2019)

Moderation effects:

- Non-causal moderation effects:
 - interaction effects (e.g. Blackwell & Olson, 2021)
 - subgroup analysis (Egami & Imai, 2019)
- Causal moderation effects:
 - for experimental data (Bansak, 2021)
 - for regression discontinuity design (Bansak & Nowacki, 2022)

What are we doing?

- Define new effect of interest
 - Balanced difference between Group Average Treatment Effects (BGATE)
- Provide assumptions that are sufficient to interpret this difference causally
 - Moderated Group Average Treatment Effect (MGATE)
- Use Double/Debiased ML to estimate the new effects of interest

Effect of interest

Setting

- N i.i.d. observations (D_i, Y_i, Z_i, X_i)
- Treatment: $D_i \in \{0, 1\}$
- Moderator: $Z_i \in \{0, 1\}$
- Potential outcomes: $\{Y_i^0, Y_i^1\}$
- Multiple treatments and moderators are possible, for simplicity only binary case in this presentation

Definition of different treatment effects

Average Treatment Effect (ATE)

$$\tau = E[Y_i^1 - Y_i^0]$$

Group Average Treatment Effect (GATE)

$$\tau(z) = E[Y_i^1 - Y_i^0 | Z_i = z]$$

Individualized Average Treatment Effect (IATE)

$$\tau(x, z) = E[Y_i^1 - Y_i^0 | Z_i = z, X_i = x]$$

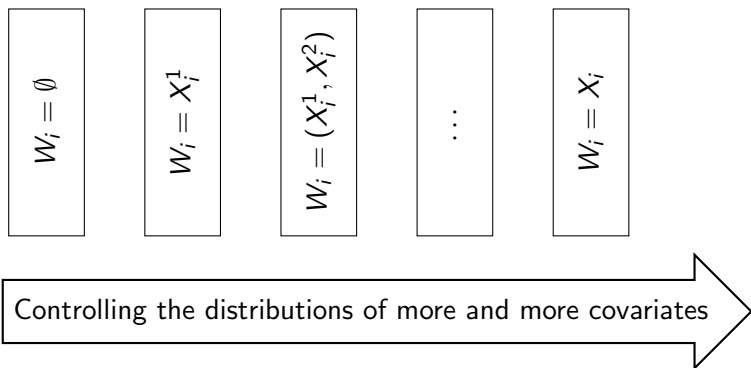
Partially balanced GATEs (BGATE)

$$\theta^B = E[E(Y_i^1 - Y_i^0 | Z_i = 1, W_i) \\ - E(Y_i^1 - Y_i^0 | Z_i = 0, W_i)]$$

with W_i being a part of X_i .

- If $W_i = \emptyset$: $\theta^B = \tau(1) - \tau(0)$ (difference between two GATEs)
- If $W_i = X_i$ and additional assumptions hold: θ^B can be interpreted as a causal moderation effect (*MGATE*)

Continuum of effects



Balanced Group Average Treatment Effect (BGATE)

We are in a **selection-on-observables** setting. The following assumptions are sufficient to identify θ^B :

- Conditional Independence Assumption (CIA)
- Common support (CS)
- Exogeneity of confounders
- Stable Unit Treatment Value Assumption (SUTVA)

Estimation: Double/debiased Machine Learning

- Framework that allows to use ML methods for causal inference (Chernozhukov et al., 2018)
- Avoids functional form assumptions
- Two ingredients to overcome bias due to use of ML methods:
 - overcoming regularization bias by using **Neyman-orthogonal score function**
 - overcoming overfitting bias by using **cross-fitting**
- Estimator is $\sqrt{n}$ -consistent and asymptotically normal and asymptotically efficient

Estimation: Double/debiased Machine Learning for ATE

- 1 Split sample into K random folds.
- 2 On $K - 1$ folds train nuisance parameters
 $\mu(d, x) = E[Y_i | X_i = x, D_i = d]$ and $p(x) = P(D_i = d | X_i = x)$
- 3 On remaining fold, predict $\hat{\mu}(d, x)$ and $\hat{p}(x)$ and calculate the pseudo outcomes $\hat{\delta}(X_i)$ by using doubly-robust score function and average it
$$\hat{\delta}(X_i) = \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i) + \frac{D_i(Y_i - \hat{\mu}(1, X_i))}{\hat{p}(x)} - \frac{(1 - D_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \hat{p}(x)}$$
- 4 Repeat steps 2 and 3 for all folds
- 5 Average ATEs

► Algorithm for ATE

Estimation: Double/debiased Machine Learning for BGATE

- 1 Split sample into K random folds.
- 2 On $K - 1$ folds train nuisance parameters $\mu(d, x)$ and $p(x)$
- 3 On remaining fold, predict $\hat{\mu}(d, x)$ and $\hat{p}(x)$ and calculate pseudo outcomes $\hat{\delta}(Z_i, X_i)$ using doubly-robust score function
- 4 **Regress them on $Z_i = z$ (e.g. age, gender) and W_i**
 - If using an ML Algo:
 - split data into J random folds, train new nuisance parameters on $J - 1$ folds and predict on remaining fold
 - use double-robust score function with moderator Z_i as “treatment”
- 5 Take difference of new pseudo outcomes and calculate BGATE
- 6 Repeat steps 2, 3, 4 and 5 for all folds
- 7 Average BGATEs

► Algorithm for BGATE

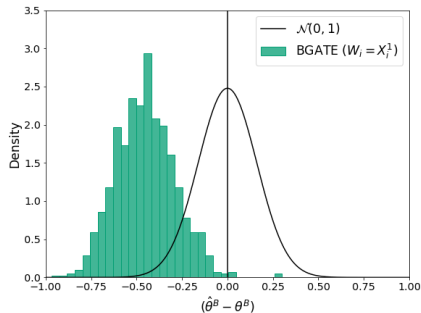
► Full DGP

- Generated data consists of a quadruple: Y_i, D_i, Z_i, X_i
- Treatment variable D_i is influenced by ten covariates $X_1, \dots, X_{10}$ and moderator Z_i in a non-linear way
- Moderator variable Z_i is influenced by two covariates X_1 and X_2 in a non-linear way
- Outcome variable Y_i is influenced by 4 covariates $X_i^1, \dots, X_i^4$, the moderator Z_i , and the treatment D_i in a non-linear way

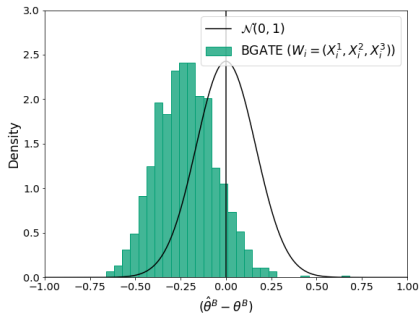
We compare the following estimators:

- Using a Radom Forest in step 2 and 3 and OLS for step 4
(non-parametric first stage, and parametric second stage)
- Using a Radom Forest in step 2, 3 and 4
(non-parametric first stage, and non-parametric second stage)

Simulation: Results by using OLS in second stage (N = 10,000)

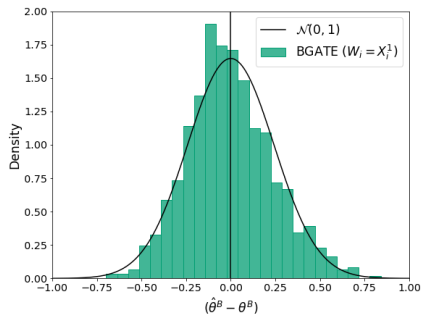


(a) BGATE with $W_i = X_i^1$

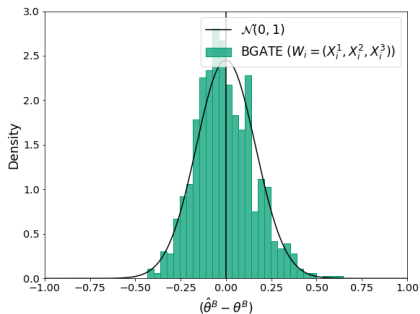


(b) BGATE with $W_i = (X_i^1, X_i^2, X_i^3)$

Simulation: Results by using Random Forest in second stage (N = 10,000)



(a) BGATE with $W_i = X_i^1$



(b) BGATE with $W_i = (X_i^1, X_i^2, X_i^3)$

Moderated Group Average Treatment Effect (MGATE)

New effect of interest

We can rewrite the special *BGATE* with $W_i = X_i$ as:

Moderated Group Average Treatment Effect (MGATE)

$$\theta^M = E \left[\left(Y_i^{1,1} - Y_i^{0,1} \right) - \left(Y_i^{1,0} - Y_i^{0,0} \right) \right]$$

- Potential outcome depend on d and z :
 $\{Y_i^{0,0}, Y_i^{0,1}, Y_i^{1,0}, Y_i^{1,1}\}$

We are still in a **selection-on-observables** setting. The following additional assumptions are sufficient to identify θ^M :

- Conditional Independence Assumption (CIA) for moderator
- Common Support (CS) for moderator
- Exogeneity of moderator
- Stable Unit Treatment Value Assumption (SUTVA) for moderator

Simulation: Z being dependent of some X 's ($N = 10,000$)

Goal: We want to estimate the MGATE (true effect of interest)

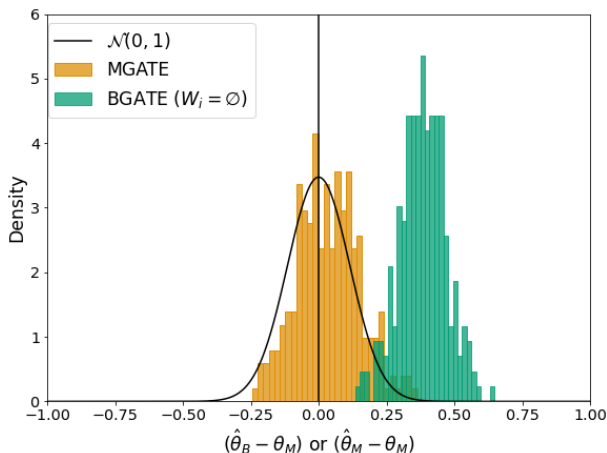


Figure: MGATE vs. BGATE (Z and some X s are dependent)

Simulation: Z and all X 's being independent ($N = 10,000$)

Goal: We want to estimate the MGATE (true effect of interest)

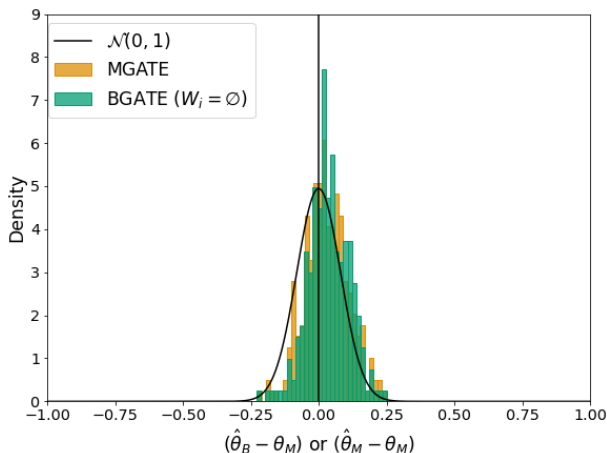
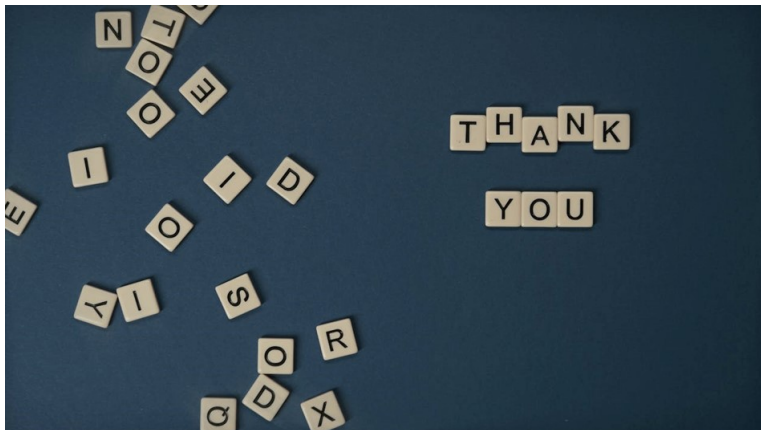


Figure: MGATE vs. BGATE (Z and all X 's are independent)

Conclusion

Conclusion

- Heterogeneous treatment effects are important
 - better understanding on how policy works
 - better understanding on which group benefits more from policy in comparison to others
- Causal Machine Learning helps us estimating them
- Interpretation is not always straightforward → be careful!
- What do we do?
 - Introduce new effect of interest (BGATE)
 - Propose estimators of BGATEs
 - Provide assumptions to interpret difference causally (MGATE)



First draft of paper available on request: Please email nora.bearth@unisg for a copy.

- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178.
- Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non-randomized moderators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1), 65–86.
- Bansak, K., & Nowacki, T. (2022). Effect heterogeneity and causal attribution in regression discontinuity designs.
- Blackwell, M., & Olson, M. P. (2021). Reducing model misspecification and bias in the estimation of interactions. *Political Analysis*, 1–20.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.

- Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants? a causal machine learning evaluation of training in belgium. *Labour Economics*, 80, 102306.
- Davis, J. M., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5), 546–550.
- Egami, N., & Imai, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, 114(526), 529-540.
- Kennedy, E. H. (2020). Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

- Knaus, M. C., Lechner, M., & Strittmatter, A. (2022). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, 57(2), 597–636.
- Lechner, M., & Mareckova, J. (2022). Modified causal forest. *arXiv preprint arXiv:2209.03744*.
- Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2), 264–289.
- Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. Forthcoming in the *Econometrics Journal*.

Appendix

DML Algorithm for ATE

Input : Data: $\{(X_i), (D_i), (Y_i)\}$

Output: ATE = $\hat{\theta}$

begin

create folds: Split sample into K random folds $(S_k)_{k=1}^K$ of observations $\{1, \dots, N\}$ with size of each fold N/K . Also define $S_k^c := \{1, \dots, N\} \setminus S_k$

for k in $\{1, \dots, K\}$ **do**

for d in $\{l, m\}$ **do**

RESPONSE FUNCTIONS:

estimate: $\mu(d, x) = E[Y_i | X_i = x, D_i = d]$ in $\{(X_i)(Y_i)\}_{i \in S_k^c, D_i = d}$

PROPENSITY SCORE:

estimate: $p(x) = P(D_i = d | X_i = x)$ in $\{(X_i), (D_i)\}_{i \in S_k^c}$

end

ATE FUNCTION:

estimate: $\hat{\theta}_k = \frac{1}{N/K} \sum_{i \in S_k} \left[\hat{\mu}(l, X_i) - \hat{\mu}(m, X_i) + \frac{l(l)(Y_i - \hat{\mu}(l, X_i))}{\hat{p}(x)} - \frac{l(m)(Y_i - \hat{\mu}(m, X_i))}{1 - \hat{p}(x)} \right]$

end

estimate effect: $\hat{\theta} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_k$

end

Algorithm 1: DOUBLE/DEBIASED ML ATE

◀ back

Estimation: Algorithm BGATE (1)

Input : Data: $\{(X_i), (Z_i), (D_i), (Y_i)\}$

Output: BGATE = $\hat{\theta}^B$

create folds: Split sample into K random folds $(S_k)_{k=1}^K$ of observations $\{1, \dots, N\}$ with size of each fold N/K . Define $S_k^c := \{1, \dots, N\} \setminus \{S_k\}$

for k in $\{1, \dots, K\}$ **do**

for d in $\{l, m\}$ **do**

 RESPONSE FUNCTIONS:

 estimate: $\mu(d, Z, X) = E[Y_i | D_i = d, X_i, Z_i]$ in $\{(X_i)(Y_i)\}_{i \in S_k^c, D_i=d}$

 PROPENSITY SCORE:

 estimate: $p(d, X, Z) = P(D_i = d | X_i = x, Z_i = z)$ in $\{(X_i), (D_i), (Z_i)\}_{i \in S_k^c}$

end

 PSEUDO-OUTCOME:

 estimate: $\hat{\delta}(Z, X) = \hat{\mu}(l, Z_i, X_i) - \hat{\mu}(m, Z_i, X_i) + \frac{l(l)(Y_i - \hat{\mu}(l, Z_i, X_i))}{\hat{p}(l, X_i, Z_i)} - \frac{l(m)(Y_i - \hat{\mu}(m, Z_i, X_i))}{\hat{p}(m, X_i, Z_i)}$ in $\{(Y_i), (X_i), (D_i), (Z_i)\}_{i \in S_k}$

end

Estimation: Algorithm BGATE (2)

create folds: Split sample S_k into J random folds $(S_j)_{j=1}^J$ of observations $\{1, \dots, J\}$ with size of each fold $N/(K \cdot J)$. Define $S_j^c := \{1, \dots, J\} \setminus \{S_j\}$

for j in $\{1, \dots, J\}$ **do**

for z in $\{v, u\}$ **do**

 PSEUDO-OUTCOME REGRESSION:

 estimate: $\zeta(z, W) = E[\hat{\delta}(Z, X)|Z_i = z, W_i]$ in $\{(W_i), (Z_i)\}_{i \in S_j^c, Z_i=z}$

 PROPENSITY SCORE:

 estimate: $p(z, W) = P(Z_i = z|W_i = w)$ in $\{(X_i), (Z_i)\}_{i \in S_j^c}$

end

BGATE FUNCTION:

EFFECT:

estimate: $\hat{\theta}_j^B = \frac{1}{N/(K \cdot J)} \sum_{i \in S_j} \left[\hat{\zeta}(v, W_i) - \hat{\zeta}(u, W_i) + \frac{I(v)(\delta(Z_i, X_i) - \hat{\zeta}(v, W_i))}{\hat{p}(v, W_i)} - \frac{I(u)(\delta(Z_i, X_i) - \hat{\zeta}(u, W_i))}{\hat{p}(u, W_i)} \right]$

STANDARD ERRORS:

estimate:

$\hat{\theta}_j^{BSE} = \frac{1}{N/(K \cdot J)} \sum_{i \in S_j} \left[\left(\hat{\zeta}(v, W_i) - \hat{\zeta}(u, W_i) + \frac{I(v)(\delta(Z_i, X_i) - \hat{\zeta}(v, W_i))}{\hat{p}(v, W_i)} - \frac{I(u)(\delta(Z_i, X_i) - \hat{\zeta}(u, W_i))}{\hat{p}(u, W_i)} \right)^2 \right]$

end

estimate effect: $\hat{\theta}^B = \frac{1}{J} \sum_{j=1}^J \hat{\theta}_j^B$

Algorithm 2: DOUBLE/DEBIASED ML FOR BGATE

p-dimensional covariates matrix X_p with $p = 6$

$$X_i \sim \mathcal{N}(0, \Sigma)$$

$$\Sigma_i = D_i^{-1} A_i^T A_i D_i$$

$$A_i \sim \mathcal{N}(0, 1)$$

$$D_i = \sqrt{\text{diag}(A_i)}$$

$$Z = \begin{cases} 1 & \text{if } \gamma \cdot \frac{X_1 + X_2}{2} + \epsilon > 0 \\ 0 & \text{else} \end{cases}$$

with $\epsilon \sim \mathcal{N}(0, 1)$ and $\gamma = \{1\}$

$$D = \begin{cases} 1 & \text{if } \frac{(X_c \times \alpha) + (0.4Z)}{4} \cdot \lambda + \epsilon > 0 \\ 0 & \text{else} \end{cases}$$

with $X_c = [X_1, X_2, X_3]$, $\alpha = [0.5, 0.3, 0.6]^T$, $\lambda = \{0.8\}$ and $\epsilon \sim \mathcal{N}(0, 1)$.

The response functions under treatment for a **linear treatment effect** are defined as:

$$\begin{aligned}\mu_{1,1}(x) &= \mu_{0,1}(x) + X_c \times \beta_1 + 0.2Z \\ \mu_{1,0}(x) &= \mu_{0,0}(x) + X_c \times \beta_0 + 0.5Z.\end{aligned}$$

with $\beta_1 = [0.7, 0.1, 0.7]^T$ and $\beta_0 = [0.2, 0.3, 0.6]^T$.

The response functions under treatment for a **nonlinear treatment effect** are defined as:

$$\begin{aligned}\mu_{1,1}(x) &= \mu_{0,1}(x) + \sin(0.7X_1 + 0.1X_2) + 0.7X_3^4 + 0.2Z \\ \mu_{1,0}(x) &= \mu_{0,0}(x) + \sin(0.2X_1 + 0.3X_2) + 0.6X_3^2 + 0.5Z.\end{aligned}$$

DGP BGATE - Potential outcomes

$$Y^{1,1} = \mu_{1,1}(x) + \epsilon_{11}$$

$$Y^{0,1} = \mu_{0,1}(x) + \epsilon_{01}$$

$$Y^{1,0} = \mu_{1,0}(x) + \epsilon_{10}$$

$$Y^{0,0} = \mu_{0,0}(x) + \epsilon_{00}$$

with errors $\epsilon_{11}, \epsilon_{01}, \epsilon_{10}, \epsilon_{00} \sim \mathcal{N}(0, 1)$. [▶ back](#)