Estimating Optimal Treatment Regimes Using Multivariate Random Forests

Boyi Guo

Department of Biostatistics

University of Alabama at Birmingham

Ruoqing Zhu

Department of Statistics

University of Illinois at Urbana-Champaign

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Outline

- Background
- Notation & Assumptions
- MedForests & MedTree
- Simulation Study
- Conclusion

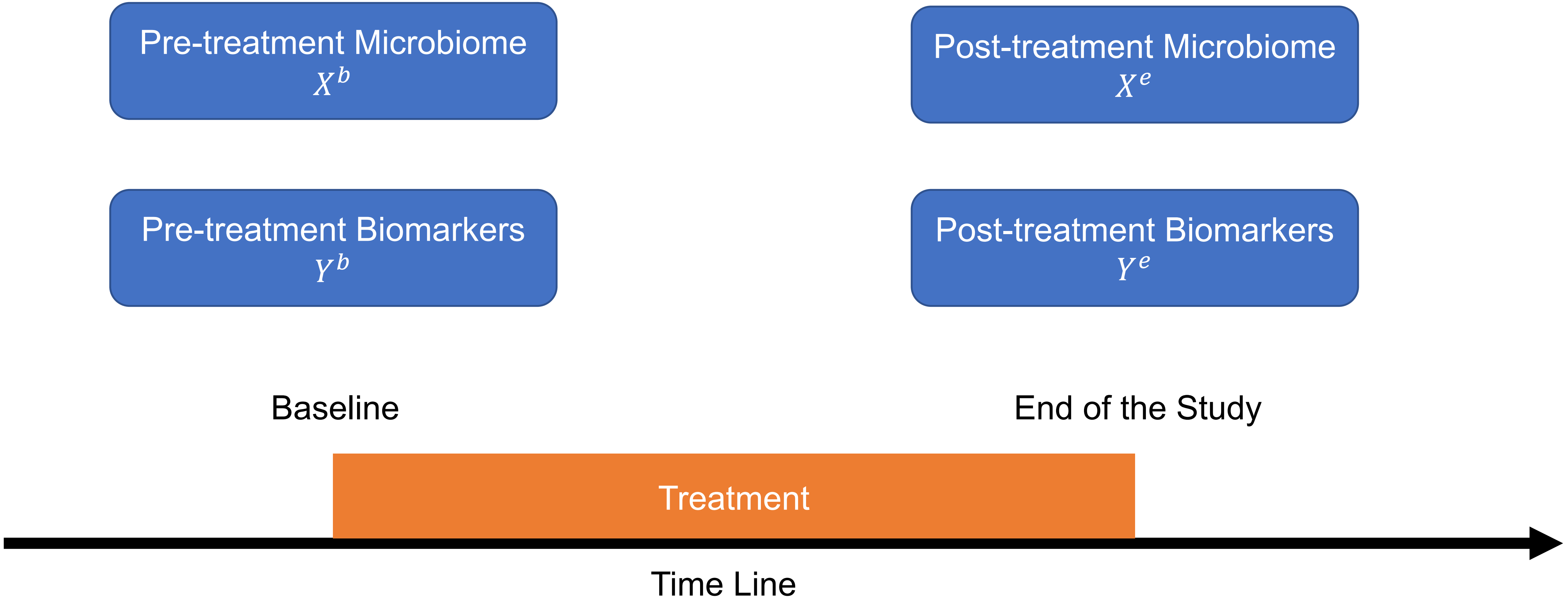
Motivation & Objectives

- A recent study (Holscher, 2018) investigated the effect of almonds on gastrointestinal microbiota and their interrelationship with human health-related biomarkers
  - Randomized controlled trial
  - Two arms of treatments
- The aim of the study is to recommend **personalized** almond diet, with respect to the microbiome composition, to improve the biomarkers

Challenges

- The treatment effect is not directly observable
- The microbiome and biomarkers are highly correlated
- Extra information collected in the study would not be fully utilized with any conventional models

Study Design

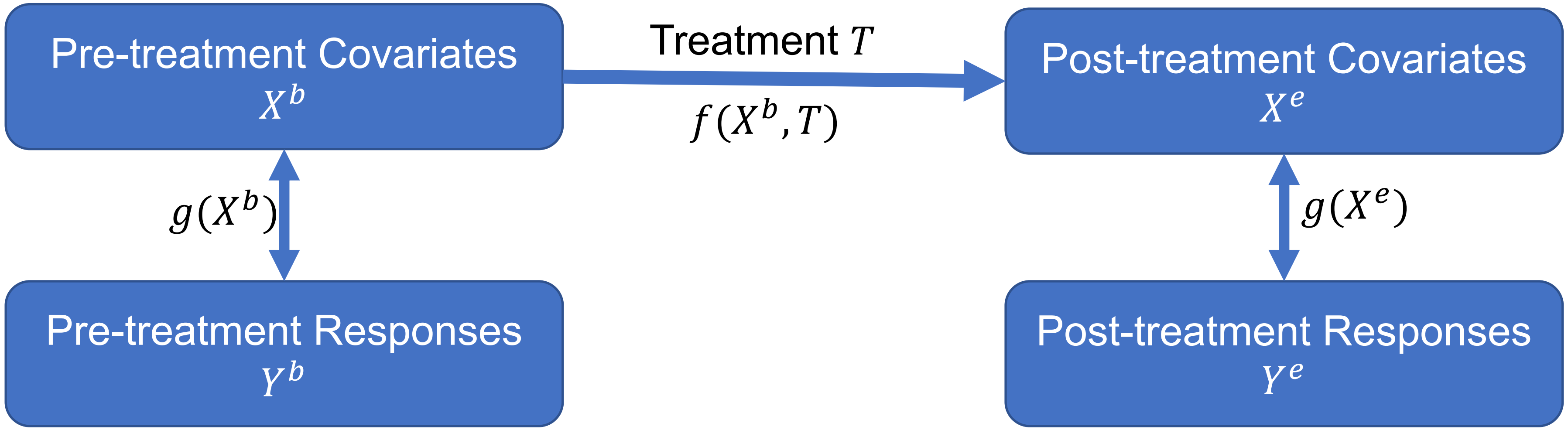


Solution

- Two-step procedure
  - Predict the post-treatment biomarkers under each arm of the treatments
  - Compare the predicted biomarkers for both treatments on a desired direction
- Predicting biomarkers conditioning on a treatment
  - A conventional solution: constructing a linear regression of the biomarkers as a function of all the interaction terms of treatment and microbiomes.
  - Multiple assumptions are not realistic: normality assumption, linear assumption, and model specification.
  - Number of variables in the model can exceed the sample size

Notation & Assumptions

- Pre- and post-treatment covariates, $X^b, X^e \in \mathbb{R}^p$
- Pre- and post-treatment responses, $Y^b, Y^e \in \mathbb{R}^q$
- Treatment, $T \in \mathcal{T} \equiv \{1, -1\}$, is independent of X^b
- Unknown treatment effect function, $f(\cdot,\cdot)$: $\mathbb{R}^p \times \mathcal{T} \to \mathbb{R}^p$
- Unknown link function from covariates to responses, g$(\cdot)$: $\mathbb{R}^q \to \mathbb{R}^q$



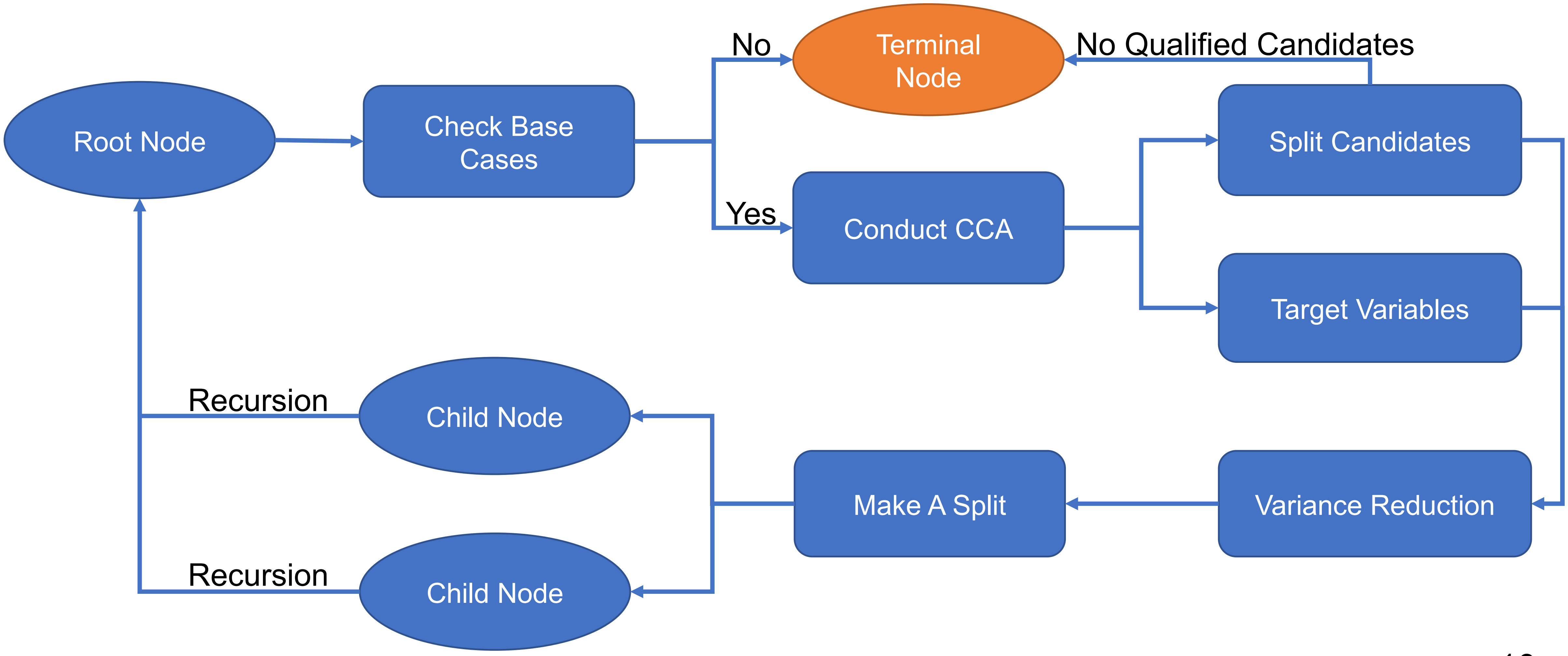
MedForests

- MedForests is an ensemble learning method that is based on random forests algorithm (Breiman, 2001)
- It consists n_{tree} distinct MedTrees, which is based on the regression tree algorithm (Breiman, 2017)
- Each MedTree recursively partitions a population, described by the pre-treatment covariates, into sub-populations where the treatment effects are similar
- When making predictions, MedForests collects the post-treatment responses that belong to the same sub-population from all MedTrees, and outputs the empirical means under each treatment arm.

Treatment Effect

- Measure of treatment effect
  - $\mathbb{E}(\Delta X|X^b,T) = \mathbb{E}(X^e - X^b|X^b,T) \; = f'(X^b,T)$
  - $\mathbb{E}(\Delta Y|X^b,T) = \mathbb{E}(Y^e - Y^b|X^b,T) \approx g(f'(X^b,T))$ under $g(\cdot)$ is linear
- Use **canonical correlation analysis (CCA)** to approximate $g(f'(\cdot,\cdot))$ while reducing dimensions
  - CCA (Hotelling, 1992) is a method for exploring the relationships between two multivariate sets of variables
  - $\underset{\rho,\beta_{T=1},\beta_{T=-1}}{arg\max}\ [corr(\rho X^b, \beta_{T=1}\Delta Y_{T=1}) + corr(\rho X^b, \beta_{T=-1}\Delta Y_{T=-1})]$
  - ρX^b is the splitting variable, and$\beta_{T=1}\Delta Y_{T=1}$,$\beta_{T=-1}\Delta Y_{T=-1}$ are the target variables to reduce variance
  - Variance Reduction:$var(\beta_{T=1}\Delta Y_{T=1}) + var(\beta_{T=-1}\Delta Y_{T=-1})$

MedTree Algorithm



Simulation Study

- 64 Data Generating Mechanisms
  - Cover various settings of dimensionality, Sample Size, Correlation Structure
  - Treatment Effect Function $f(X,T) \in \{Linear,\ Circle,\ Box\}$
  - Link Function $g(X) \in \{\beta X\ , \beta X^2\}$ where $X^2\ =\ x_{ij}^2$ for all pairs of i, j
  - 200 iterations
- Models Compared
  - $\mathcal{L}1$-Penalized Least Square (Qian & Murphy, 2011)
  - GUIDE (Loh, He & Man, 2015)
- Performance Metric
  - Averaged recommendation error rate

Simulation Study Result (p=20, q=6)

		Uncorrelated				Correlated			
	Setting	**MedTree**	**MedForest**	L_1**PLS**	**GUIDE**	**MedTree**	**MedForest**	L_1**PLS**	**GUIDE**
$N = 400$	Linear	0.168 (0.048)	0.078 (0.025)	**0.038 (0.010)**	0.262 (0.035)	0.256 (0.081)	0.129 (0.047)	**0.040 (0.011)**	0.328 (0.076)
	Circle	**0.416 (0.022)**	0.419 (0.020)	0.434 (0.020)	0.428 (0.017)	0.366 (0.066)	**0.309 (0.092)**	0.587 (0.041)	0.352 (0.112)
	Box	0.222 (0.034)	**0.188 (0.012)**	0.189 (0.012)	0.202 (0.046)	0.267 (0.048)	**0.204 (0.021)**	0.258 (0.014)	0.321 (0.081)
	Square	0.436 (0.033)	**0.345 (0.034)**	0.485 (0.020)	0.501 (0.028)	0.411 (0.049)	**0.316 (0.054)**	0.456 (0.042)	0.480 (0.058)
$N = 800$	Linear	0.155 (0.044)	0.062 (0.021)	**0.026 (0.008)**	0.229 (0.037)	0.229 (0.057)	0.105 (0.033)	**0.029 (0.008)**	0.278 (0.052)
	Circle	**0.388 (0.024)**	0.391 (0.033)	0.428 (0.016)	0.420 (0.031)	0.275 (0.049)	**0.183 (0.045)**	0.572 (0.024)	0.286 (0.086)
	Box	0.215 (0.023)	**0.188 (0.013)**	0.189 (0.013)	0.199 (0.029)	0.237 (0.030)	**0.177 (0.016)**	0.259 (0.015)	0.269 (0.064)
	Square	0.401 (0.027)	**0.288 (0.026)**	0.482 (0.021)	0.497 (0.030)	0.368 (0.037)	**0.240 (0.032)**	0.473 (0.050)	0.468 (0.047)

Conclusion

- MedForests & MedTree
  - Learning models that require few assumptions
  - Incorporate extra information to improve accuracy of estimation
  - Utilize correlation structures to reduce dimensions.
  - Outperform traditional models when treatment effect function is complicated
- Future Directions
  - Tuning Parameters
  - Interpretability
    - variable importance measure

Acknowledgements

- University of Illinois at Urbana Champaign
  - Dr. Hannah D. Holscher
  - Loretta S. Auvil
  - Michael E. Welge
  - Colleen B. Bushell
- Beltsville Human Nutrition Research Center
  - Dr. David J. Baer
  - Dr. Janet A. Novotny
- University of Alabama at Birmingham
  - The Graduate Student Government

GRADUATE STUDENT GOVERNMENT

GSG

UNIVERSITY OF ALABAMA AT BIRMINGHAM®

References

- Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
- Breiman, L. (2017). *Classification and regression trees*. Routledge.
- Holscher, H., Taylor, A., Swanson, K., Novotny, J., & Baer, D. (2018). Almond Consumption and Processing Affects the Composition of the Gastrointestinal Microbiota of Healthy Adult Men and Women: A Randomized Controlled Trial. *Nutrients, 10*(2), 126. doi:10.3390/nu10020126
- Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics* (pp. 162-190). Springer, New York, NY.
- Loh, W. Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, *34*(11), 1818-1833.
- Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics, 39*(2), 1180-1210. doi:10.1214/10-aos864

Appendix: Treatment Effect

- $\mathbb{E}\left(\Delta Y \middle| X^b, T\right) = \mathbb{E}(Y^e - Y^b | X^b, T) = \mathbb{E}\left(g(X^e) - g\left(X^b\right) \middle| X^b, T\right)$
  - If $g(\cdot)$ is linear, $\mathbb{E}\left(\Delta Y \middle| X^b, T\right) = \mathbb{E}\left(g(\Delta X) \middle| X^b, T\right) = g(\mathbb{E}\left(\Delta X \middle| X^b, T\right) = g(f'(X^b, T))$

Appendix: Simulation Setting

- 64 Data Generating Mechanisms
  - Dimensionality $(p, q) \in \{(10,3),\ (20,6)\}$
  - Training Sample Size $N \in \{400,\ 800\}$, Testing Sample Size $N_{test} = 1000$
  - Treatment Effect Function $f(X, T) \in \{Linear,\ Circle,\ Box\}$
  - Link Function $g(X) \in \{\beta X\ , \beta X^2\}$ where $X^2 = x_{ij}^2$ for all pairs of i, j
  - $X^b \sim MVN(0, \Sigma)$ where $\Sigma \in \{\sigma^2\mathbb{I}, AR(0.8)\}$
  - $T \sim\ Bernolli(0.5)$
  - $Y = g(X)\ +\ MVN(0, \sigma_Y^2\mathbb{I}), X^e = f(X^b, T) + MVN(0, \sigma_x^2\ \mathbb{I})$
  - 200 iterations