CAUSAL INFERENCE: BAYESIAN CAUSAL FOREST (BCF), BAYESIAN ADDITIVE REGRESSION TREES (BART) AND MULTILEVEL TEST(MLT)

Authors: Huaxing Zeng, Jing Xu, Xiaoke Song, Yaqi Liu, Yuyan Fan
Affiliation: Brown Data Science Institute
Github repo: https://github.com/xiaoke-song/data2020_final_project.git

Contribution of Each Member:
Xiaoke was responsible for designing and implementing the simulation framework for BCF; Jing led the evaluation and visualization of model performance across both simulation and empirical analyses; Yuyan handled the simulation components for BART and MLT; Huaxing managed data preprocessing and ran BCF on the real-world NMES dataset; Yaqi conducted BART and MLT experiments on the same dataset.

Introduction:

This project aims to estimate heterogeneous treatment effects (HTEs) under confounding, using Bayesian Causal Forests (BCF), Bayesian Additive Regression Trees (BART), and multilevel models. Our statistical goal is to compare these methods in terms of accuracy and uncertainty when estimating conditional average treatment effects (ATEs) from observational data.

Our work builds on Hahn et al. (2020), who introduced BCF to reduce bias caused by regularization in causal models. By separating treatment and prognostic effects and incorporating the estimated propensity score, BCF improves HTE estimation. We extend their framework by adding multilevel models and testing all methods on real and simulated data to assess robustness and practical utility.

We replicate the data-generating process from Section 6.1 of Hahn et al. (2017) to compare these models, evaluating performance using RMSE for CATE (combined with its Confidence Interval and Standard Deviation). We then apply the same methods to the 1987 National Medical Expenditure Survey (NMES) dataset to estimate the causal effect of heavy smoking on medical expenditures.

Methods:

BCF vs. BART vs. MLT

Bayesian Causal Forest
Bayesian Additive Regression Tree
Multilevel Test

Simulated Data:

- Simulations use two sample sizes (250, 500) across 4 settings: linear/nonlinear $\mu(x)$ and homogeneous/heterogeneous $\tau(x)$.
- 5 covariates affect both outcomes and treatment. Propensity score depends on $\mu(x)$ and X1, simulating strong confounding.
- Outcome is $Y=\mu(x)+\tau(x)Z+\varepsilon$, allowing comparison of model bias, RMSE, and coverage.

NMES Data (1987):

- Outcome: log of annual medical expenditures.
- Treatment: heavy smoking (≥17 pack-years).
- Covariates:
  - Age, gender, race, marital status, education, income, region
  - Smoking start age, years since quitting, seatbelt use

Evaluation & Comparision:

- RMSE(CATE/ATE):

We use the RMSE of the estimated Conditional and Average Treatment Effects to assess model accuracy, as it reflects both bias and variance in estimation. Lower RMSE indicates better predictive performance.

- Confidence Interval:

We compute 95% confidence intervals for the RMSE of CATE to quantify the uncertainty around model performance.

- Standard Deviation:

Standard deviation of RMSE across simulations is used to assess the variability of model performance. Smaller standard deviations indicate more consistent model behavior across different data realizations.

Simulation & Results:

For Simulated Data

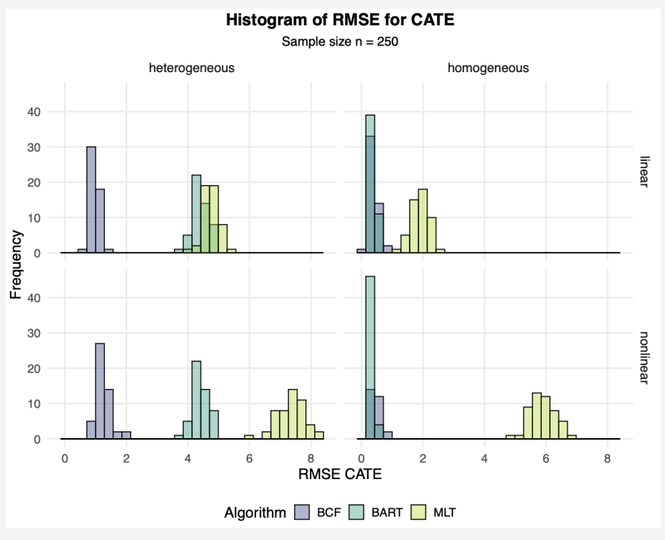

Figure 1. Histogram of RMSE for CATE Estimates Across Methods and Data-Generating Scenarios (n = 250)

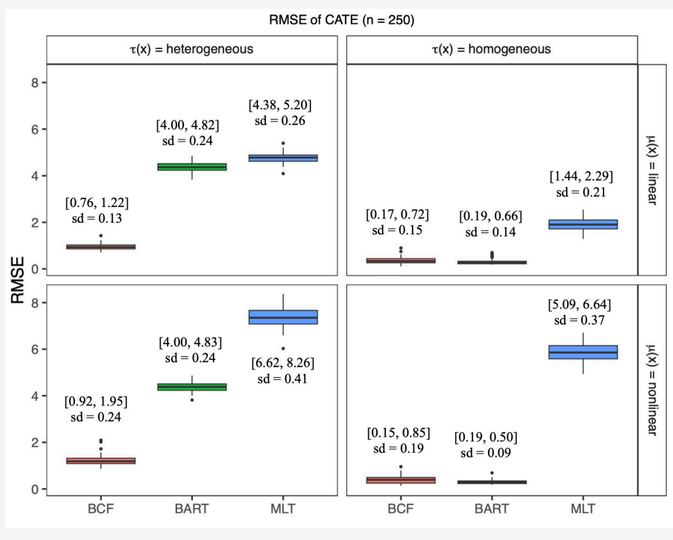

Figure 2. Boxplot of RMSE for CATE Estimates by Model and Data-Generating Scenario (n = 250)

- **BCF: Achieves the lowest RMSE across all settings, showing strong robustness to both treatment effect heterogeneity and nonlinearity in outcomes.**
- **BART: Performs reasonably well, but RMSE increases in nonlinear or heterogeneous cases, indicating sensitivity to complexity.**
- **Multilevel (MLT): Performs worst, with high RMSE under heterogeneity and nonlinear response surfaces, reflecting its limited flexibility in capturing individual-level treatment variation.**

For NMES Data

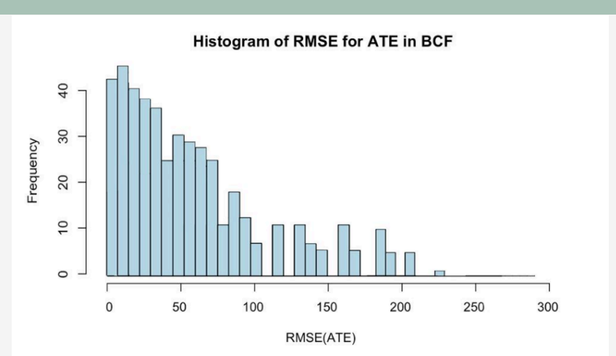

Figure 3: Histogram of RMSE for ATE Estimates Using BCF

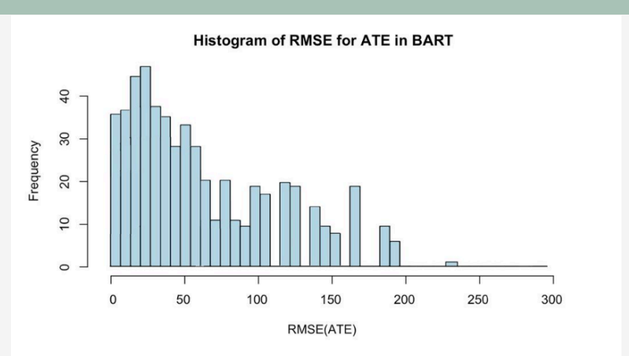

Figure 4: Histogram of RMSE for ATE Estimates Using BART

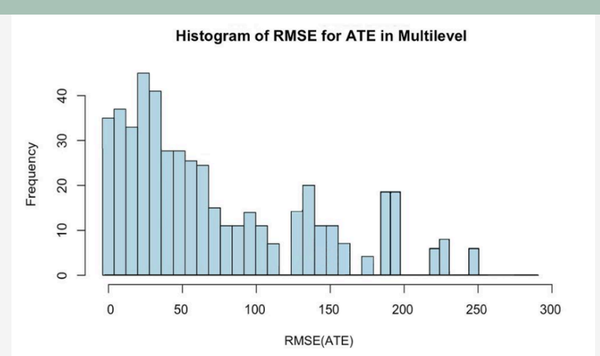

Figure 5: Histogram of RMSE for ATE Estimates Using Multilevel Model

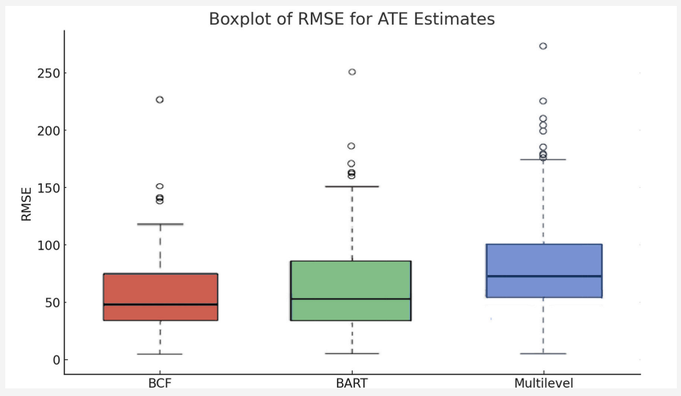

Figure 6: Boxplot of RMSE for ATE Estimates Across Models(BCF, BART, Multilevel)

- **BCF: demonstrates the lowest median RMSE and the narrowest interquartile range, highlighting its stability and accuracy.**
- **BART: shows moderately higher error dispersion, with more frequent outliers beyond 150.**
- **MLT: exhibits the highest variability, a larger median RMSE, and a heavy tail of extreme values, suggesting it is more prone to large estimation errors.**

Conclusion:

Our results demonstrate that BCF provide the most accurate and stable estimates for treatment effects in both simulation and real-world settings, outperforming BART and MLT. This highlights the strength of Bayesian methods for causal inference, which is consistent with Hahn et al. (2017).

In the context of estimating heterogeneous treatment effects, BCF's explicit modeling of treatment and prognostic components—along with propensity scores—makes it especially well-suited for settings with confounding and complex interactions, like those in our simulations and the NMES data. Its strong performance suggests it can better capture individual-level variation in treatment effects.

Challenges:

1. high computational cost due to repeated MCMC sampling in BART/BCF.
2. Median imputation is straightforward but may distort variance estimates and weaken predictive performance.

Improvements:

1. using cross-validation to tune model parameters for better generalization
2. collecting time series features to capture longitudinal patterns in treatment response.

References:

1. Hahn et al. (2017), Bayesian regression tree models for causal inference, arXiv:1706.09523
2. NMES Data. National Medical Expenditure Survey. Available from: https://CRAN.R-project.org/package=causaldrf
3. Imai, K., & van Dyk, D.A. (2004). Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. Journal of the American Statistical Association, 99(467).
4. National Center for Health Services Research and Health Care Technology Assessment. NATIONAL MEDICAL EXPENDITURE SURVEY, 1987: INSTITUTIONAL POPULATION COMPONENT. Rockville, MD: Westat, Inc. [producer], 1987. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1990. doi:10.3886/ICPSR09280.v1
5. Bryer, Jason M. "TriMatch: An R Package for Propensity Score Matching of Non-binary Treatments." The R User Conference, useR! 2013 July 10-12 2013 University of Castilla-La Mancha, Albacete, Spain. Vol. 10. No. 30. 2013.