International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Propensity Score Analysis in Observational Studies with Confounders and Missing Data

Margaret A. Batocael[1], Bernadette F. Tubo[2]

[1,2]Department of Mathematics and Statistics, MSU-IIT, Iligan City, Philippines

Abstract

This study explores the use of propensity score matching to reduce bias in estimating treatment effects from observational data. Specifically, it evaluates the performance of logistic regression and machine learning-based methods for propensity score estimation under conditions involving missing data and complex confounding structures. Simulation studies were conducted using both complete and imputed datasets across varying levels of missingness, unmeasured confounding, and nonlinearity in the true propensity score. Logistic regression (LR), generalized boosting models (GBM), and Bayesian additive regression trees (BART) were compared based on estimation accuracy and covariate balance. Performance was assessed using root mean square error (RMSE) mean absolute error (MAE), R-squared, absolute standardized mean differences (ASMD), and Kolmogorov–Smirnov (KS) statistics. The results highlight trade-offs in model robustness, particularly between predictive accuracy and covariate balance, offering practical insights for selecting appropriate propensity score models in complex observational settings.

Keywords: Propensity Score, Observational Study, Missing Data, Unmeasured Confounding

1. Introduction

Observational studies are essential for investigating causal effects when randomization is not feasible due to ethical or practical constraints. First introduced by Rosenbaum and Rubin (1983), propensity score (PS) methods are widely used to simulate randomization by balancing covariates across treatment groups [6]. Traditionally estimated using logistic regression, PS can now also be derived using machine learning (ML) techniques. These advanced methods offer greater flexibility and robustness to model misspecification. Among the four main PS-based approaches, namely: matching, stratification, weighting, and covariate adjustment, matching is often favored for its simplicity and effectiveness in achieving covariate balance.

Despite the growing use of PS matching, challenges remain, particularly, in the presence of unmeasured confounders and missing data. Unmeasured confounders are those variables that influence both the treatment and outcome but are not accounted for. On the other hand, missing data has mechanisms, to name: missing completely at random (MCAR) where the missing values are not related to the observed and unobserved, missing at random (MAR) where the missing values are related only to the observed, and missing not at random (MNAR) where missing values are related to the unobserved data itself [5]. Observational studies frequently encounter these issues, complicating causal inference.

Following the recommendation of Kim et al. (2023), this study investigates the PS estimation and matching by comparing traditional and ML-based estimation models under various conditions [2]. These include the presence or absence of unmeasured confounders, different missing data mechanisms, and varying levels of missingness. The main goal is to provide deeper insights into the effectiveness and

reliability of PS estimation and matching with high-performing PS models in the literature in supporting causal inference in real-world, data-challenged environments.

2. Methodology

This study employed a simulation-based approach to investigate the performance of PS estimation and matching in the presence of unmeasured confounders and missing data.

Data Generation: Observational data were generated and patterned in the simulation of Sturmer et al. (2010) to mimic a cohort study design with a binary treatment variable (T) (see Table 1 in the Appendix for details) [7]. The data comprised ten observed covariates $(X_1, \dots, X_{10})$ (three binary and seven continuous with standard normal distribution, $N(0,1)$). The predicted probability of the intended treatment was calculated based on the ten measured covariates using a logistic model, and was used to assign two unmeasured binary covariates (X_{11}, X_{12}) representing frailty-like conditions [7]. The probability of actual treatment was recalculated based on the ten measured and two unmeasured covariates. Generally, four conditions were introduced in the true propensity score model:

- Condition 1: without interaction and quadratic terms, with unmeasured confounders

$$P(T \mid X_1, \dots, X_{12}) = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + \dots + a_{10} X_{10} + a_{11} X_{11} + a_{12} X_{12})}} \quad (1)$$

- Condition 2: without interaction and quadratic terms, without unmeasured confounders

$$P(T \mid X_1, \dots, X_{10}) = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + \dots + a_{10} X_{10})}} \quad (2)$$

- Condition 3: with interaction and quadratic terms, with unmeasured confounders

$$P(T \mid X_1, \dots, X_{12}) = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + \dots + a_{12} X_{12} + c(X_8 X_9) + d(X_{10})^2)}} \quad (3)$$

- Condition 4: with interaction and quadratic terms, without unmeasured confounders

$$P(T \mid X_1, \dots, X_{10}) = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + \dots + a_{10} X_{10} + c(X_8 X_9) + d(X_{10})^2)}} \quad (4)$$

Treatment assignment was then generated using equation (5), given by

$$T = \{1, \ U \leq P(T \mid X)\ 0, \quad otherwise \quad (5)$$

Each condition has 1,000 generated datasets each containing 1,000 samples. Moreover, the parameters and values covered in the data generation are presented in Table 1 of the Appendix.

Missing Data Simulation and Imputation: Multivariate MCAR and MAR data were introduced at proportions of 2%, 5%, 10%, and 15% across all 1,000 datasets under each of the four specified conditions using *delete_MCAR* and *delete_MAR* censoring functions in the R software. These mechanisms were determined using a detection framework described in Figure 1. This ensures that the generated missing data constituted true MCAR and MAR. Predictive Mean Matching (PMM) was then applied to the incomplete datasets to handle missing values. PMM was chosen because of its strong imputation performance [3]. Since PMM can only impute values based on observed data, this study considered only

MCAR and MAR mechanisms, as MNAR involves unobserved data and cannot be appropriately addressed by PMM.

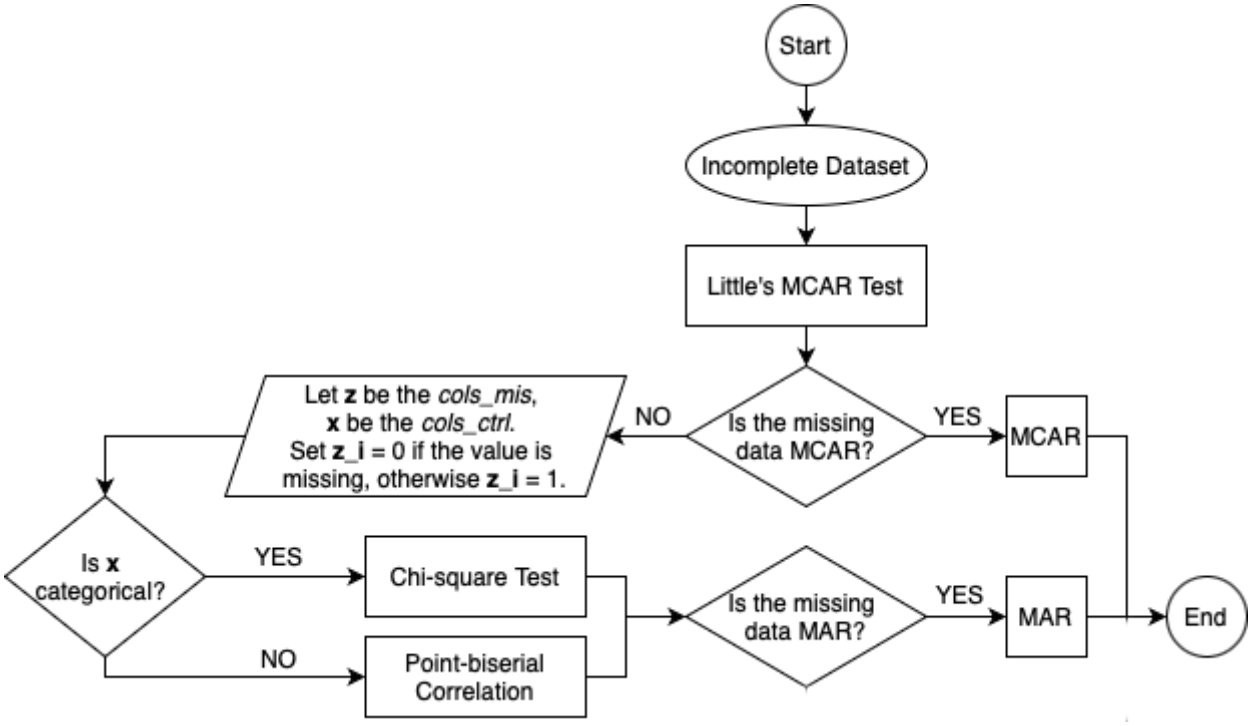

Figure 1: Framework for Detecting a Type of Missing Data Mechanism

PS Estimation and Matching: Propensity scores were estimated on both complete and imputed datasets using three techniques: logistic regression (LR), generalized boosting models (GBM), and Bayesian additive regression trees (BART). These models were selected to represent conventional and machine learning approaches (most and least used models [4]), allowing performance comparison under various data complexities. The most used GBM was fitted using a number of iterations at 10,000, a shrinkage factor of 0.001, bag of 1.0, and a depth of interaction of 5.0. On the other hand, BART was fitted using the default hyperparameters. Predictive accuracy of the models was evaluated using root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination, R^2. The models' significant difference was tested based on R^2 using Kruskal-Wallis and was further assessed by Post-Hoc analysis. Furthermore, different models estimate PS differently, which affects their ability to create comparable treatment and control groups. After estimation, nearest neighbor caliper matching without replacement was applied as a matching algorithm. The caliper was 0.2 of the standard deviation of the logit of the PS [1]. Covariate balance post-matching was assessed using the absolute standardized mean differences (ASMD) and Kolmogorov-Smirnov (KS) statistics. Estimation and matching were performed using the *matchit* function in R.

3. Results and Discussion

The average performance was assessed based on 1,000 datasets under each of the four specified conditions introduced in the true propensity score model.

Complete Cases: When main effects are only introduced in the true propensity score model, LR demonstrated the best performance, followed by BART and GBM, regardless of whether unmeasured

confounding was present. However, it is shown in Figure 2 (right) what happens to LR when used to estimate propensity scores under such conditions where there are non-additivity and nonlinearity components in the true propensity score model. In contrast, machine models adapted better to these complex conditions, with BART consistently achieving strong performance. While most methods exhibit significant differences (refer to Table 2 in the Appendix), BART and GBM do not significantly differ when unmeasured confounding is present (p-value = 0.7726259 > 0.05), suggesting their comparable performance under this complex data condition.

Another key observation is that the presence of unmeasured confounding leads to decreased accuracy and increased error levels. This suggests that unmeasured confounding complicates the relationships between covariates, treatment, and outcome. As a result, propensity score models may struggle to accurately estimate the true propensity scores.

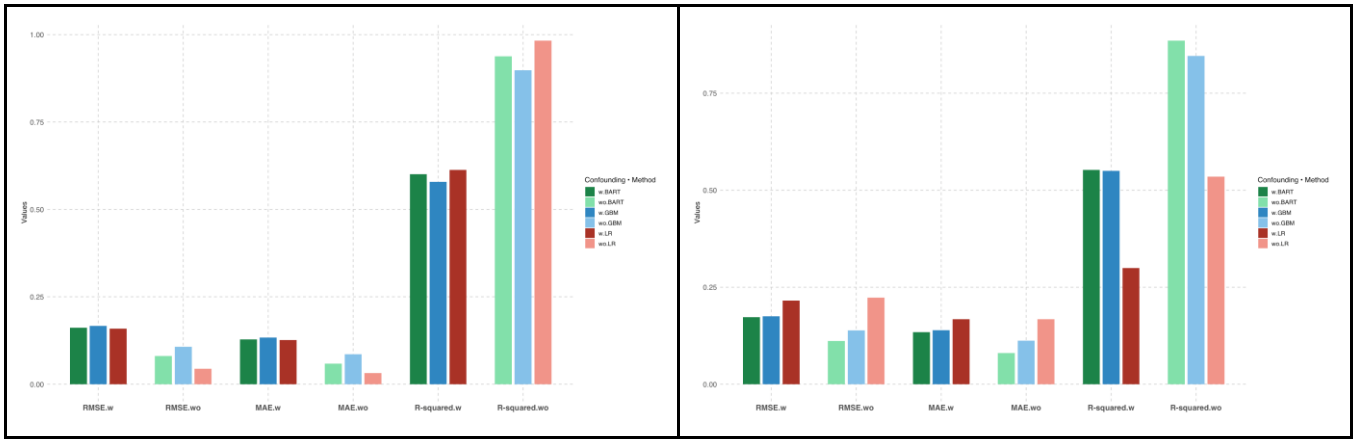

Figure 2: Average Model Performance Across 1,000 Simulated Datasets (Left: Conditions 1 and 2, Right: Conditions 3 and 4)

The results of matching when main effects were only introduced in the true PS model with and without unmeasured confounding are presented in Figure 3 (A and B). In both ASMD and KS measures, matched BART-estimated PS was able to balance all covariates while LR and GBM struggle balancing certain covariates, suggesting potential limitations in their ability to achieve adequate covariate balance. The same result can be observed when nonlinear and non-additive components were introduced (Figures 3C and 3D).

Imputed MCAR and MAR Cases: In the case where there were missing values but imputed, all models showed degraded performance with increasing missingness (refer to Figures 4 and 5). In most imputed MCAR cases, BART performs best consistently. It is comparable with GBM at 2% with unmeasured confounding and at 15% without unmeasured confounding, under complex conditions. The same result can be observed in cases with imputed MAR data. BART and GBM are only comparable at 2% and 5% with unmeasured confounding under complex conditions (refer to Tables 3 and 4 in the Appendix).

In the model's covariate balancing ability under imputed MCAR, it is seen that in all cases, matching from BART-estimated PS consistently achieves good covariate balance with and without unmeasured confounding, under both simple and complex conditions. The same results show under imputed MAR. LR and GBM struggle to balance some covariates in most levels of imputed MCAR and MAR data under all specified conditions of the true PS.

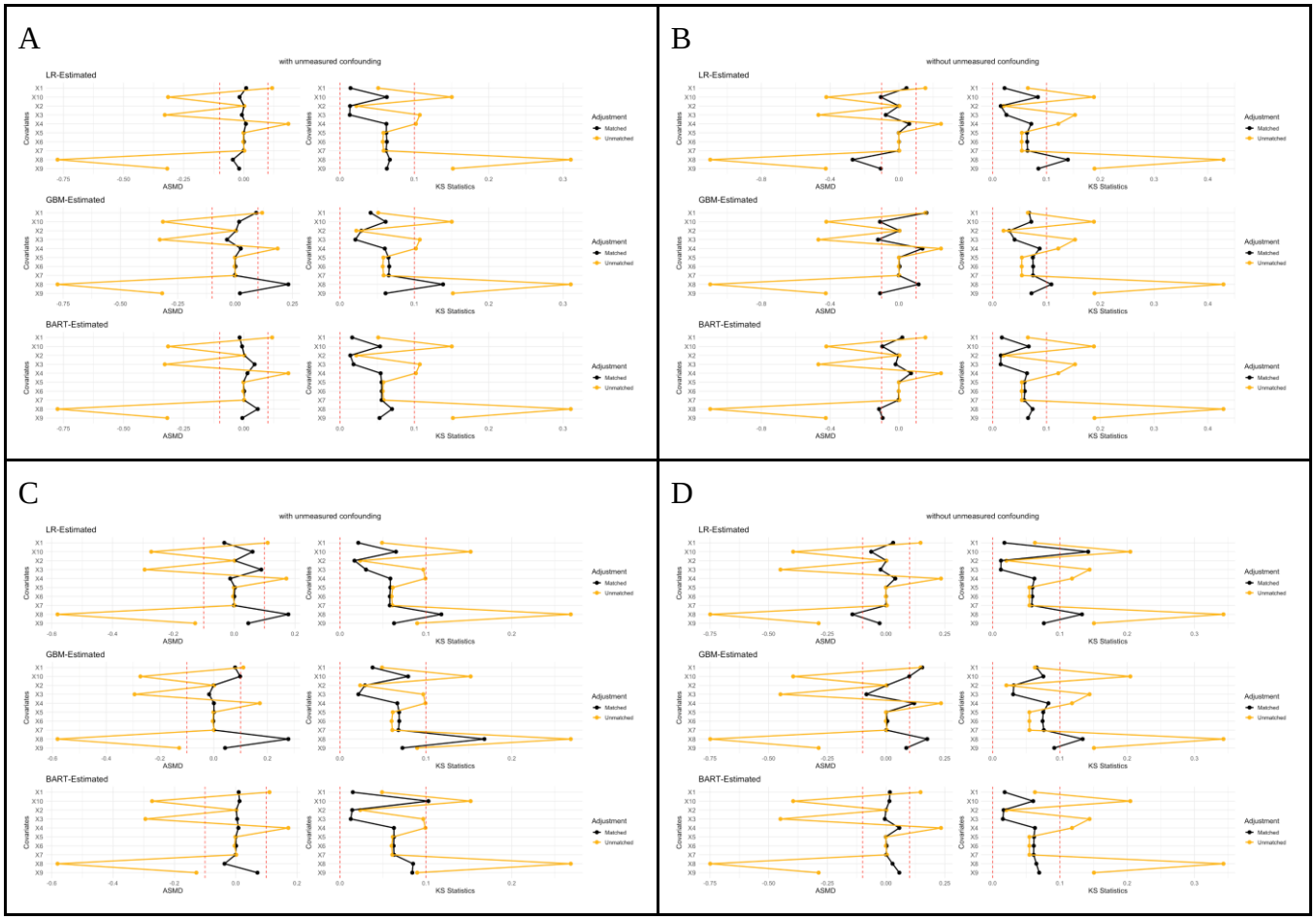

Figure 3: Average Balance Measures Across 1,000 Simulated Datasets (A: Condition 1, B: Conditions 2, C: Condition 3, and D: Condition 4)

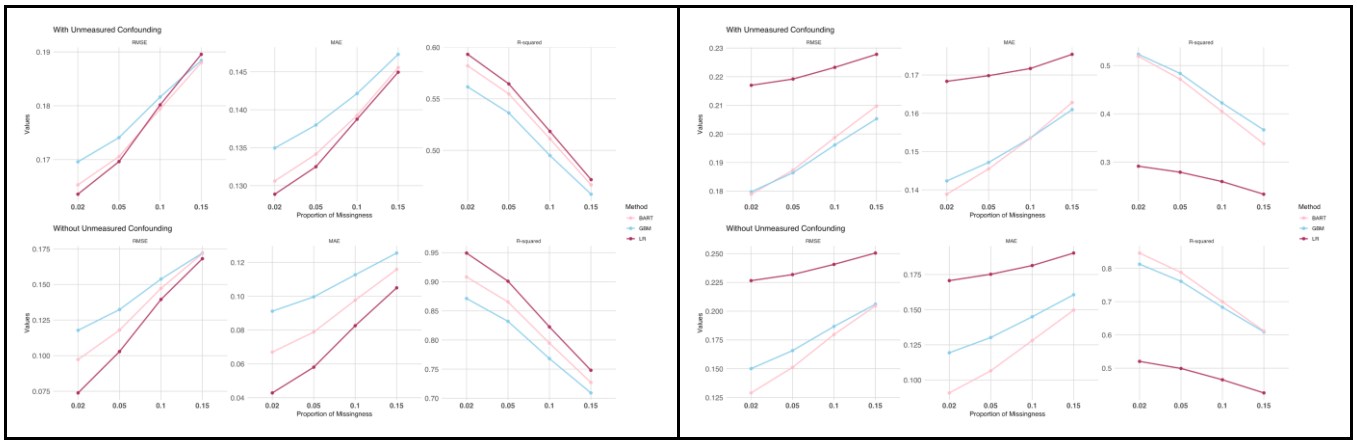

Figure 4: Average Model Performance Across Different Proportions of Imputed MCAR Data (Left: Conditions 1 and 2, Right: Conditions 3 and 4)

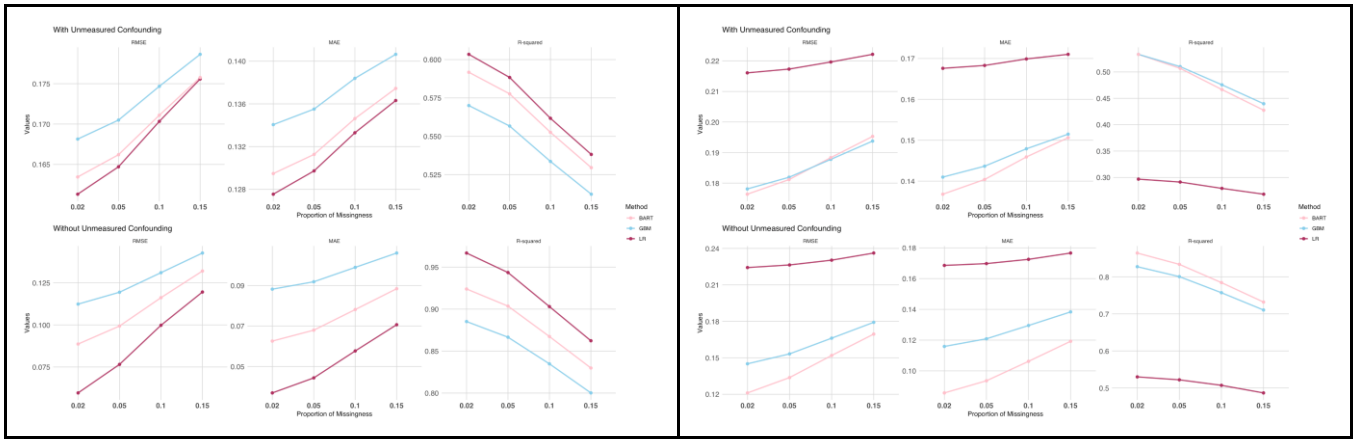

Figure 5: Average Model Performance Across Different Proportions of Imputed MAR Data (Left: Conditions 1 and 2, Right: Conditions 3 and 4)

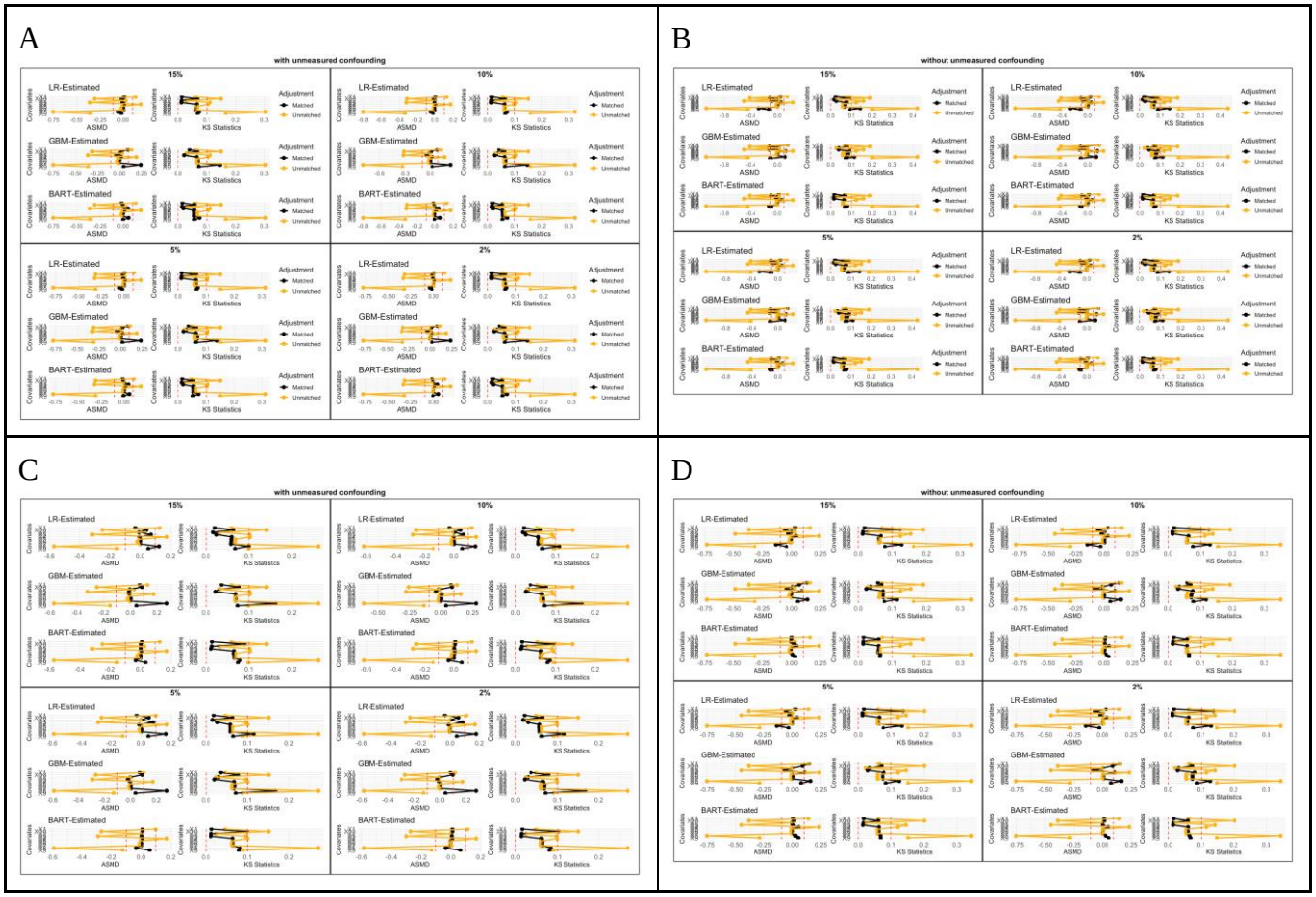

Figure 6: Average Balance Measures Across Different Proportions of MCAR Data (A: Condition 1, B: Conditions 2, C: Condition 3, and D: Condition 4)

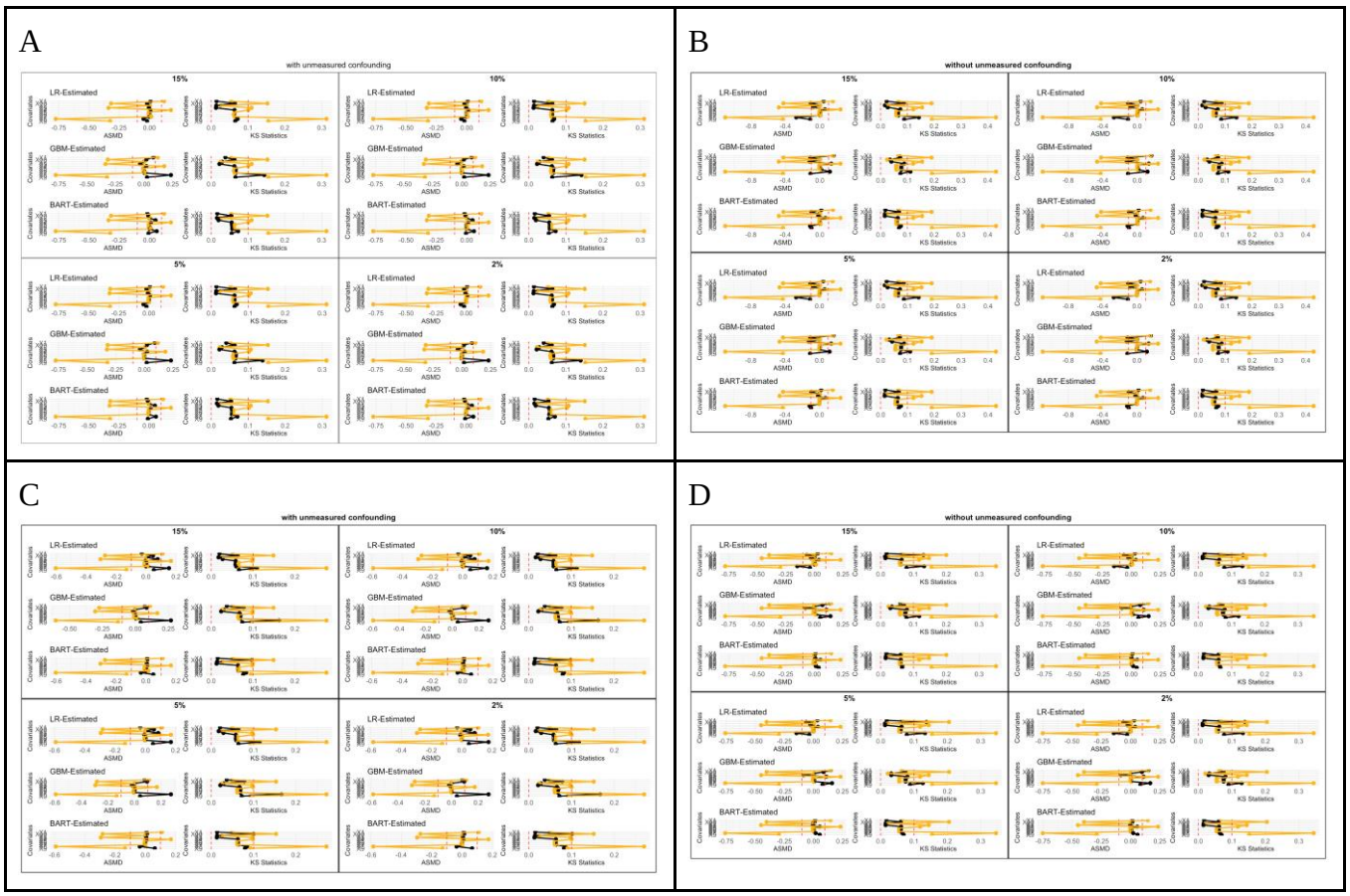

Figure 7: Average Balance Measures Across Different Proportions of MAR Data (A: Condition 1, B: Conditions 2, C: Condition 3, and D: Condition 4)

4. Conclusion

This study found that logistic regression (LR) exhibits notable limitations in propensity score analysis, especially under complex data structures or when model specification is difficult. In contrast, machine learning approaches, particularly Bayesian Additive Regression Trees (BART) demonstrated superior performance in both predictive accuracy and covariate balance, even in the presence of missing data and potential unmeasured confounding. While the generalized boosting model (GBM) is commonly used as a nonparametric machine learning method, it struggled to consistently balance covariates across simulation settings. BART, though less frequently applied in practice, emerged as a robust and flexible alternative, offering substantial improvements in balance and model robustness when estimating propensity scores. These findings provide the importance of selecting propensity score methods that prioritize both predictive accuracy and covariate balance. Given its adaptability to complex, nonlinear data structures, BART offers a powerful alternative for propensity score estimation and matching in observational research, outperforming other nonparametric methods that are often favored for their minimal reliance on strict data assumptions.

References

1. Benedetto, U., Head, S. J., Angelini, G. D., & Blackstone, E. H. (2018). Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, *53*, 1112-1117.

2. Kim, S., Lee, J., & Jung, K. (2023). Propensity Score Estimation Using Neural Networks: A Comparison of DNN, CNN, and Logistic Regression. *Research square*.
3. Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics, 42*, 371-404.
4. Leite, W., Zhang, H., Collier, Z., Chawla, K., Kong, L., Lee, Y., & Leite, W. L. (2024). Machine Learning for Propensity Score Estimation: A Systematic Review and Reporting Guidelines. *OSF Preprints*.
5. Little, R. J., Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.
6. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.
7. Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American journal of epidemiology, 172*, 843-854.