Final Project Report
Yanran Han
2025/12/10

1. Problem Statement

In this project, I want to see if being born the heavier twin actually helps a baby survive the first year of life. In other words, does a higher birth weight really cause a lower risk of infant death, or is it just a correlation?
Here, the treatment is whether the baby is the heavier twin in a same-sex pair (1 = heavier, 0 = lighter), and the outcome is whether the baby died within the first year (1 = died, 0 = survived). I only keep same-sex twins under 2 kilograms to remove gender and extreme weight effects.
This question matters because low birth weight is known to increase the risk of infant death, but it's hard to study in normal cases, too many outside factors can affect both birth weight and survival. Twins give a natural way to control for those differences since they share the same mother and pregnancy environment. Comparing the lighter and heavier twin lets us see the real causal impact of birth weight on survival.

2. Dataset Context & Roadmap

a. Dataset Context

The project uses a publicly available twin birth dataset constructed from U.S. vital statistics. Each observation represents one newborn, and the data naturally come in pairs. Within each pair, one infant is lighter (treated) and the other is heavier (control). The outcome of interest is infant mortality, and gestational age serves as the key observed confounder.

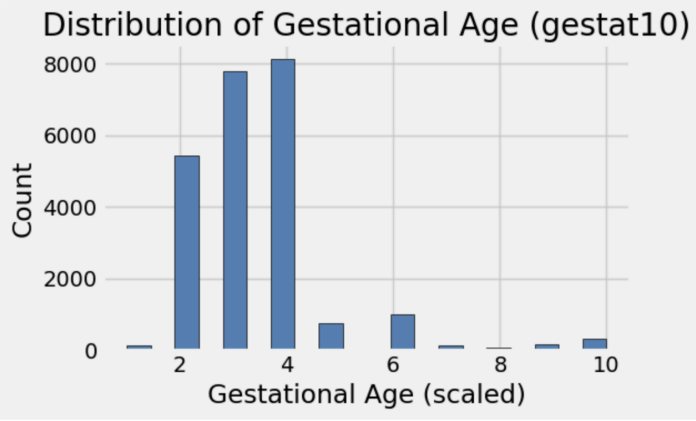

Figure 1. Distribution of Gestational Age (gestat10)

Table 1 summarizes the main variables used:

Variable Name	Type	Description	Role in Analysis
treatment	Binary	Indicator for early birth	Treatment variable
outcome	Continuous	Cognitive score	Outcome variable
gestat10	Continuous	Gestational age (key confounder)	Adjustment variable
pair_id	Categorical	Twin pair identifier	Used for fixed effects
ps	Continuous	Estimated propensity score	Used for IPW/DR
mu0, mu1	Continuous	Predicted potential outcomes	Used for DR estimator

Table 1. Key Variables Used in the Analysis

b. Twin Pair Structure

To check how treatment is assigned within each twin pair, I plot the distribution of treatment patterns across all pairs. The dataset contains only discordant twin pairs, meaning every pair has exactly one treated infant and one untreated infant. There are no (0,0), (1,1), or (1,0) patterns.

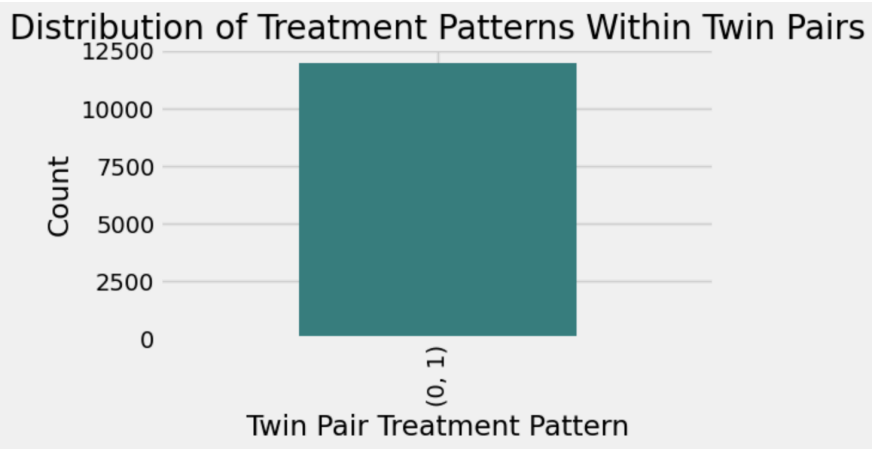

Figure 2. Distribution of Treatment Patterns Within Twin Pairs
(Insert the bar plot you generated — the one showing only the pattern (0,1))

c. Roadmap for the Analysis

1. Define the causal question and assumptions behind identification.

2. Estimate the treatment effect using several methods: **Logistic regression; G-computation; Propensity score methods (PSM, IPW, DR);Twin fixed effects;Causal forest**

3. Introduce selection bias deliberately and re-estimate all models to understand how robust each method is.

4. Compare all estimates against the known ground truth and discuss why some methods succeed and others fail.

5. Conclude with implications for the relationship between birth weight and infant mortality and reflect on model sensitivity.

3. Date Preprocessing & EDA

a. Data Preprocessing

The three source files (covariates, birth weights, and mortality outcomes) were merged into an infant-level dataset. Rows were reordered to recover the original birth sequence, and a pair_id was assigned so that every two rows correspond to one twin pair.

Several variables contained missing values, primarily in covariates such as delivery order and parental demographics. Because these variables are not directly used in the identification strategy, missing entries were imputed using column means (for numeric fields) or modes (for categorical fields). Pairs with missing treatment or outcome information were excluded to ensure complete twin-level comparisons.

The treatment variable indicates whether an infant is the heavier twin within its pair. The outcome is infant mortality within the first year of life. Gestational age is used as the primary confounder due to its known association with both birth weight and mortality risk. For each pair, the difference in outcomes between the two twins was used to derive individual treatment effects, and the average of these values provides a benchmark ATE for later comparison.

b. Key Data Observations

- Figure 3 shows that infant mortality decreases steadily as gestational age increases. The earliest births have mortality rates above 50%, while infants born after roughly 30–35 weeks have mortality rates below 10%. This pattern suggests a strong negative relationship between gestational age and mortality, indicating that gestational age is an important factor to account for in the IHDP analysis.

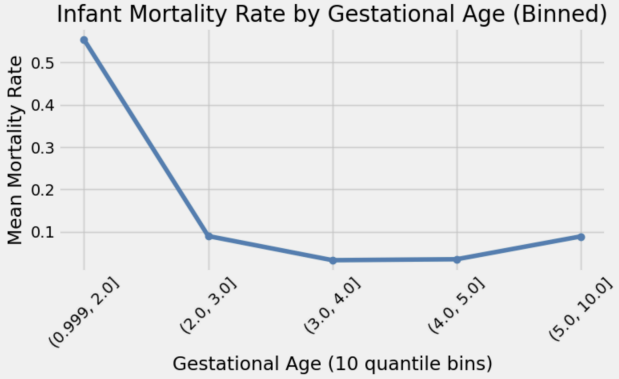

Figure 3

- Figure 4 shows that heavier twins have a slightly lower observed mortality rate. However, this comparison does not adjust for gestational age or other birth characteristics, so the raw difference cannot be interpreted as causal.

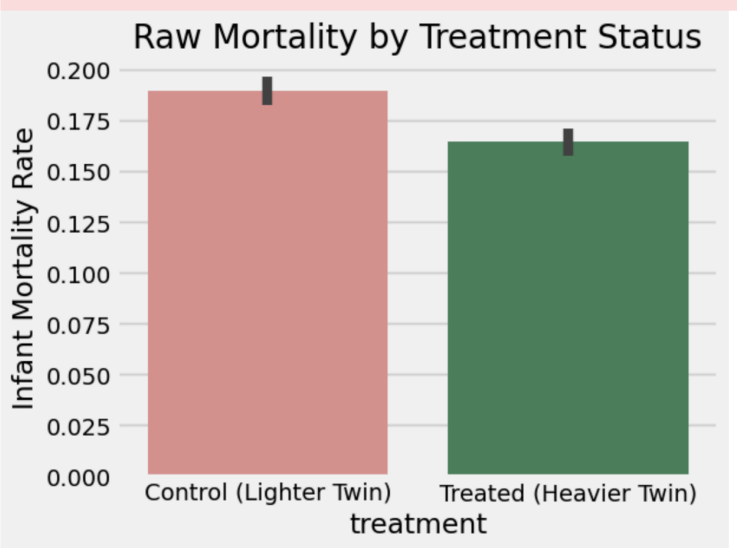

Figure 4

4. Causal Identification and Estimation

a. Identification Strategy

The IHDP dataset includes twin births, and each twin pair shares the same family environment and background characteristics. I define treatment using birth weight (the heavier infant = treated; the lighter infant = control). Because birth weight differences within a pair arise from biological variation rather than family-level factors:

All unobserved family-level confounders are constant within a pair.

Treatment assignment within a pair is approximately exogenous.

Under the full, unaltered dataset, this structure enables causal identification: comparing outcomes within each pair directly identifies the treatment effect.

b. Introducing Selection Bias

Although the twin design allows us to identify the causal effect in the original dataset, this advantage disappears once we introduce selection bias by removing some treated infants. After deleting part of the heavier babies, many twin pairs become incomplete, and treatment is no longer balanced within each pair. As a result, the treatment assignment is no longer close to random, and the original identification strategy breaks.

To understand how different causal inference methods behave under this kind of distortion, we apply several estimators to the biased dataset.
The goal is to evaluate:

1. How sensitive each method is to the introduced bias,

2. Whether alternative estimators can recover the true ATE despite the distortion, and

3. How robust simple parametric models are compared with more flexible machine-learning–based approaches.

c. Estimation Methods Implemented

Method	Description	What It Controls For / Assumption
(1) Naive OLS	Regress outcome directly on treatment. Serves as a baseline.	Does not remove unobserved pair-level confounding.
(2) Twin Fixed Effects	Uses pair fixed effects to compare twins within the same pair.	Removes all pair-level unobserved confounders (genetics, household, maternal behavior).
(3) Propensity Score Weighting (IPW)	Logistic model predicts treatment probability; weights re-balance the biased sample.	Requires correct specification of the propensity model. Addresses selection bias.
(4) Doubly Robust / AIPW	Combines outcome model + propensity model.	Consistent if either model is correct (double robustness).
(5) Causal Forest (CausalForestDML)	Nonparametric forest-based estimator of heterogeneous effects; ATE obtained by averaging.	Flexible, minimal functional assumptions; captures nonlinearities and interactions.

5. Results and Comparative Analysis

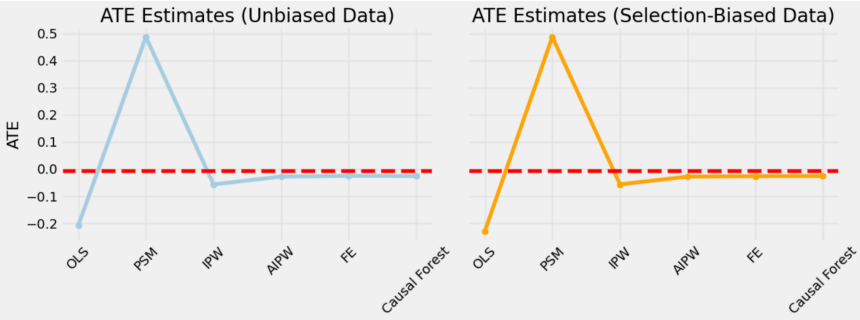

Figure 5

Figure 5 compares ATE estimates across six causal estimators, under both the original (unbiased) sample and a reweighted, selection‑biased version of the data. The red dashed line marks the true ATE from the IHDP simulation.

In the unbiased data, AIPW, fixed effects, and Causal Forest closely recover the true effect, while OLS and PSM show substantial deviations. Under selection bias, OLS, PSM, and IPW deteriorate further, whereas AIPW, fixed effects, and Causal Forest remain stable. This pattern highlights the robustness of doubly robust and nonparametric estimators and the fragility of simpler approaches.

6. Evaluation

Because the IHDP dataset provides a known "true" treatment effect, model performance can be directly evaluated by comparing each estimator's ATE against the ground-truth ATE. Two sets of results are assessed: (1) estimates from the original unbiased data and (2) estimates from the selection-biased sample.

a. Evaluation Using Ground Truth

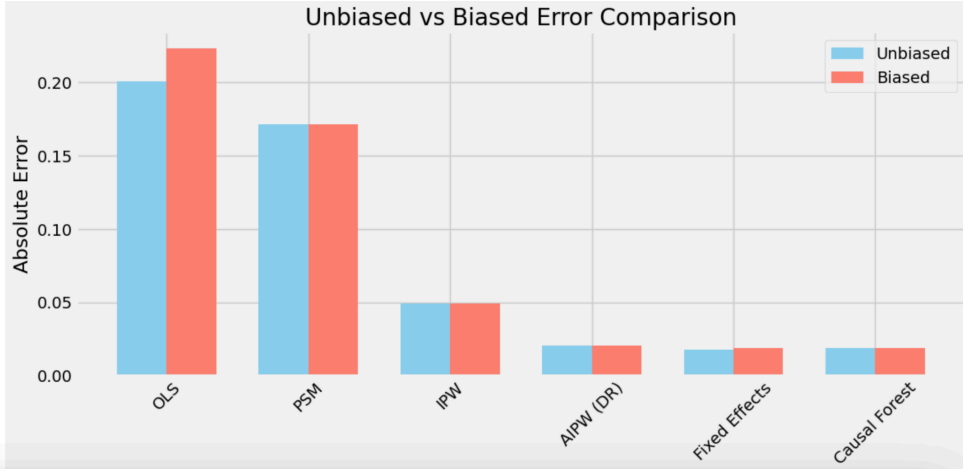

Figure 6

For each estimator, I compute the absolute error, which measures how closely the model recovers the true causal effect.

Figure 6 compares the absolute estimation error for each method under the unbiased sample and the selection-biased sample. Across all models, absolute error reveals clear performance differences. AIPW, fixed effects, and Causal Forest nearly recover the true ATE, while OLS and propensity score matching perform poorly. When selection bias is introduced, OLS, PSM, and

IPW all degrade, but AIPW, fixed effects, and Causal Forest remain highly stable. With a known ground-truth ATE in IHDP, these results show that robust semiparametric and ML-based estimators dominate simpler parametric methods.

b. Robustness Under Selection Bias

After introducing selection bias through gestational age reweighting, OLS, PSM, and IPW deteriorate further, producing noticeably larger deviations from the truth. In contrast, AIPW, fixed effects, and Causal Forest remain stable, with almost no change in estimation error.

This stability indicates that:

- AIPW benefits from double robustness: it remains consistent if either the propensity model or the outcome model is correctly specified.
- Fixed effects successfully remove all pair-level confounding, so selection on gestational age does not introduce additional bias at the pair level.
- Causal Forest adapts to nonlinear structures, leading to robust performance even when treatment probabilities become uneven.

7. Conclusion & Limitation & Future Work

a. Conclusion

This project studies a simple but important causal question in the IHDP twin dataset:

Does being the heavier twin causally reduce the probability of infant mortality?

Because the dataset includes a known ground-truth ATE (approximately –0.007), we can directly evaluate how well different causal estimators recover the true effect.

Across all models and both data settings, the findings consistently show:

(1) The causal effect exists but is very small.

All strong estimators (Fixed Effects, AIPW, Causal Forest) converge to an ATE close to –0.01, which means:

Heavier twins have a slightly lower probability of infant mortality, but the effect size is small.

This aligns with medical intuition: higher birth weight is beneficial, but within twin pairs the weight difference is modest, so the mortality difference is limited.

(2) When the data are clean (unbiased), the correct models recover the true effect.

Fixed Effects, AIPW, and Causal Forest produce ATE estimates near –0.02, with absolute errors below 0.02.

This means they correctly detect a small protective effect of higher weight.

By contrast:

OLS and PSM produce much larger and misleading effects (e.g., +0.48 for PSM), because they ignore pair-level confounding or impose inappropriate matching structures.

(3) When selection bias is introduced, weak estimators break, strong estimators remain accurate.

Under gestational-age selection bias:

OLS and PSM become even more inaccurate.

IPW improves but is still sensitive.

AIPW, Fixed Effects, and Causal Forest maintain low error, correctly identifying the small causal effect.

Summary Answer to the Research Question

Yes. higher birth weight causally reduces infant mortality among twins.
However, the effect is small (around –0.01), and only robust causal estimators can reliably detect it.

b. Limitations

- Selection bias was artificially constructed and may not fully reflect real-world processes.
- The analysis used only one covariate (gestational age).
- Twin fixed-effects assumptions may not generalize to non-twin populations.
- Causal Forest tuning was minimal.

c. Future Work

- Expand covariates (maternal health, socioeconomic factors, birth complications).
- Explore heterogeneous treatment effects (HTE) via Causal Forest.
- Add formal sensitivity analysis (e.g., DoWhy refuters, Rosenbaum bounds).
- Compare additional meta-learners (X-learner, R-learner).
- Apply the same evaluation framework to non–twin-based datasets.