Estimating the Causal Effect of Early Childhood Intervention on Cognitive Development

Huanlin Dai
ADSP 32029 – Autumn 2025
Instructor: Jeong-Yoon Lee

1. Causal Question and Motivation

Research Question

What is the causal effect of participation in an intensive early childhood intervention program on cognitive test scores among low-birth-weight, premature infants?

Motivation

Early childhood is a crucial period of life where interventions can significantly impact developmental disparities. Premature infants with lower birthweight are at higher risk of cognitive and developmental delays, further motivating targeted intervention. However, determining the causal effect of such programs proves difficult due to potential systematic differences between families who enroll in intervention programs and those who do not.

Understanding early intervention's effects can help shape better policy decisions and possibly funding/scaling of similar programs. This analysis aims to contribute to that understanding by applying causal inference methods to quantify treatment effects while addressing confounding factors.

Dataset, Context, and Scope

This analysis uses the **IHDP (Infant Health and Development Program)** dataset, specifically the version included in the *PyWhy / DoWhy* package. Although IHDP was originally a randomized experiment, this version introduces nonrandom treatment assignment to simulate an observational study setting. This makes the dataset especially valuable for teaching and evaluating causal inference tools. As such, the scope of this analysis is limited to studying how well each method recovers the effect of a treatment on the outcome.

Roadmap

This report proceeds as follows:

1. **Exploratory Data Analysis** describes the dataset, main variables, and preprocessing steps.
2. **Causal Identification and Estimation** outlines assumptions and implements three causal inference methods:
    - Propensity Score Overlap Weighting
    - Augmented Inverse Propensity Weighting (AIPW)
    - Meta-Learners
3. **Results and Comparative Analysis** summarize estimated effects and compare method outputs.
4. **Evaluation** uses robustness checks and ground-truth ATE from IHDP.
5. **Conclusion and Future Work** synthesizes insights and discusses limitations.
6. **Link to the Code Repository** is provided at the end.

2. Exploratory Data Analysis

2.1 Data Source

The dataset is loaded following the DoWhy IHDP example. The IHDP contains **747 infants** described by several variables/metrics.

Because the PyWhy version simulates selection bias, treatment assignment is no longer randomized, making it appropriate for observational causal inference. However, the covariates have all been anonymized/rescaled to prevent attempts to trace individuals included in the data.

2.2 Variable Description

- **Treatment (T):** Participation in the early childhood intervention program (0/1)
- **Outcome (Y):** Stanford–Binet IQ score at 36 months (y_factual)
- **Counterfactual Outcome (Y_C):** Simulated counterfactual IQ scores (y_cfactual)
- **True potential outcomes (μ_0, μ_1):** Simulated potential outcomes for each individual

- **Covariates (X):** Anonymized maternal, familial, and birth-related characteristics (e.g., birth weight, gestational age, maternal education, socioeconomic variables)

These variables are potential confounders because they may influence both treatment enrollment and child cognitive outcomes.

2.3 Initial EDA Highlights

- Mean treatment rate: 18.6% treated (139/747)
- Average IQ score:
    - Treated: 6.432418 (std 1.108928)
    - Control: 2.411297 (std 1.595977)
- Raw difference (naïve estimate): 4.021 points
  This difference is *not* causal due to confounding.
- Propensity scores show moderate imbalance between groups.

Figure 1. Visualization of class imbalance between control and treatment groups

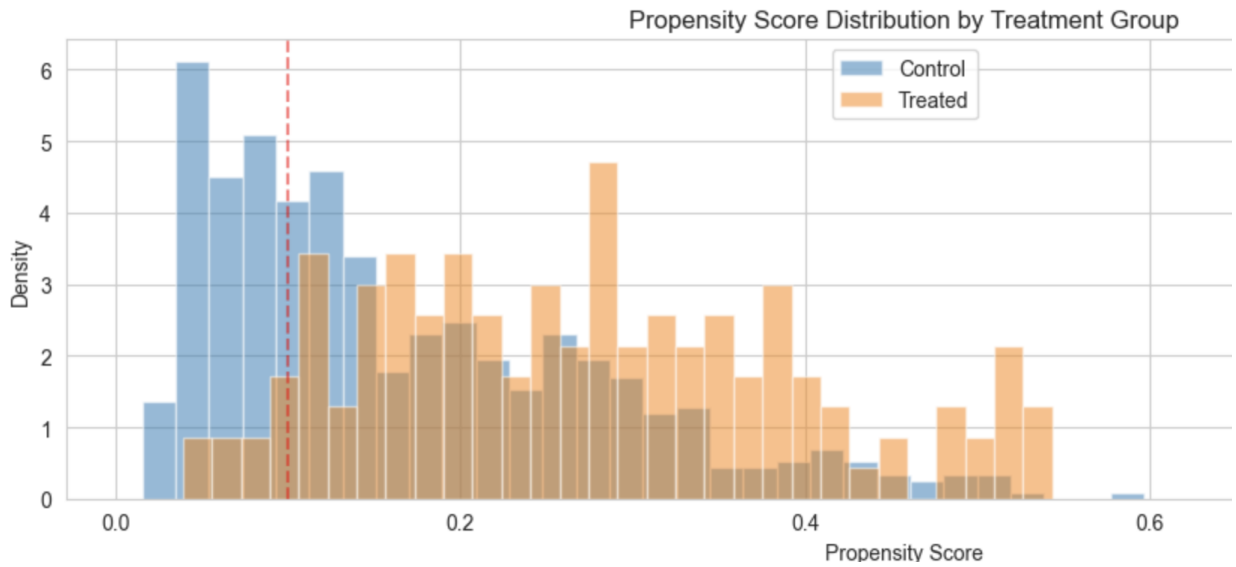


Figure 2. Propensity Score Distributions

3. Causal Identification and Estimation

This section formalizes the causal estimand, explicitly states the identification assumptions, and documents the causal inference methods used: Propensity Score Overlap Weighting, Augmented Inverse Propensity Weighting (AIPW), and Meta-Learner approaches for estimating treatment effect heterogeneity.

3.1 Target Estimands

3.1.1 Average Treatment Effect (ATE)

The primary estimand is the Average Treatment Effect:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

where

- Y(1) = potential outcome (IQ score at 36 months) if treated
- Y(0) = potential outcome if untreated

The challenge is to identify this estimand using observational (non-randomized) treatment assignment from IHDP.

3.1.2 Conditional Average Treatment Effect (CATE)

While the ATE provides a population-level summary, it can mask substantial heterogeneity in treatment effects across subgroups. To understand **who benefits most** from early intervention, we also estimate the Conditional Average Treatment Effect:

$$\mathrm{CATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

This individual-level estimand allows us to:

1. Identify subpopulations with larger or smaller treatment effects
2. Inform targeting and resource allocation decisions
3. Test whether treatment effects vary systematically with baseline characteristics

3.2 Identification Assumptions

To identify both ATE and CATE from observational data, the following assumptions are required.

(1) Stable Unit Treatment Value Assumption (SUTVA)

SUTVA consists of two parts:

1. **No interference**: One infant's treatment status does not affect another infant's cognitive score.
    - This is reasonable because early intervention participation at the family level does not plausibly affect other families.
2. **Consistency**: The observed outcome equals the potential outcome corresponding to the treatment received.
    - Standard assumption for causal inference.

(2) Conditional Independence Assumption (CIA)

$$(Y(0), Y(1)) \perp T \mid X$$

Given the rich set of IHDP covariates (birth weight, SES, maternal behavior, etc.), we assume that the remaining unobserved confounding is not strong enough to overturn results. The possibility of unmeasured confounding is discussed in Evaluation.

(3) Positivity (Common Support)

$$0 < P(T = 1 \mid X) < 1 \text{ for all individuals}$$

This requires that for every covariate profile, both treatment and control observations exist. Because IHDP includes meaningful imbalance, positivity must be checked and addressed. Overlap weighting is especially attractive because it automatically downweighs regions of poor overlap.

3.3 Estimation Methods

3.3.1 ATE Estimation via Weighting Methods

Propensity Score Overlap Weighting

Overlap weighting assigns weights that emphasize the region of covariate space where treated and control units are most similar, automatically handling positivity violations in the tails of the propensity score distribution. Weights are defined below.

$$w_i = \begin{cases} 1 - e(X_i) & \text{if } T_i = 1 \\ e(X_i) & \text{if } T_i = 0 \end{cases}$$

The propensity score is denoted as follows.

$$e(X_i) = P(T = 1 \mid X_i)$$

Augmented Inverse Propensity Weighting (AIPW)

AIPW combines outcome regression with propensity score weighting to achieve double robustness—the estimator is consistent if either the outcome model or the propensity score model is correctly specified.

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i(Y_i - \hat{\mu}_1(X_i))}{e(X_i)} - \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - e(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right]$$

where $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ are estimated outcome models for treated and control units, respectively.

3.3.2 CATE Estimation via Meta-Learners

To estimate treatment effect heterogeneity, we use four meta-learner approaches that leverage machine learning methods to estimate $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$:

S-Learner (Single Model)

Estimates a single outcome model with treatment as a covariate:

$$\hat{\mu}(x, t) = \mathbb{E}[Y \mid X = x, T = t]$$

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

This is a simple and efficient learner, but may underestimate heterogeneity if the treatment effect is small relative to baseline risk. We will eventually see that the opposite is true, and the S-Learner actually overestimates heterogeneity.

T-Learner (Two Models)

Estimates separate outcome models for treated and control groups:

$$\hat{\mu}_1(x) = \mathbb{E}[Y \mid X = x, T = 1], \quad \hat{\mu}_0(x) = \mathbb{E}[Y \mid X = x, T = 0]$$

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

More flexible than S-Learner, but can be unstable in regions with few observations in either treatment arm.

X-Learner (Cross-validated Imputation)

A three-stage procedure designed to handle imbalanced treatment assignments:

1. Estimate outcome models for treated and control groups separately (as in T-Learner)
2. Impute individual treatment effects:
    - For treated individuals: estimated effect = observed outcome - predicted control outcome

    - For control individuals: estimated effect = predicted treated outcome - observed outcome
3. Estimate CATE models on the imputed effects and combine via propensity score weighting

Performs well when one treatment group is much smaller, as in IHDP (approximately 25% treated).

R-Learner (Robinson Transformation)

Uses a residual-on-residual regression approach with double robustness properties:

1. Estimate the expected outcome given covariates and the propensity score (probability of treatment) via cross-fitting
2. Compute residuals:
    - Outcome residuals = observed outcome minus predicted outcome
    - Treatment residuals = actual treatment minus predicted treatment probability
3. Estimate CATE by regressing outcome residuals on covariates, weighted by the squared treatment residuals

The R-Learner is particularly robust to model misspecification and handles confounding effectively by explicitly controlling for both outcome and treatment propensities.

Base Learners

For all meta-learners, we use gradient boosting (GradientBoostingRegressor) as the base learner due to its strong performance on tabular data and ability to capture non-linear relationships and interactions without extensive feature engineering.

Evaluation

Because the IHDP simulation dataset includes both factual and counterfactual outcomes μ_0 and μ_1, we can directly compute the true CATE and evaluate meta-learner performance using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 (proportion of CATE variance explained)
- Correlation between predicted and true CATE

This allows validation of heterogeneous treatment effect estimates, which is typically impossible in real-world observational studies.

4. Results and Comparative Analysis

This section reports the estimated causal effects from the multiple methods used above. Results are compared to assess robustness and identify subpopulations that benefit most from early intervention.

4.1 Average Treatment Effect (ATE) Estimates

Table 1 summarizes the estimated ATE of early childhood intervention on IQ at age 36 months.

Table 1. Estimated ATEs Across Methods

Method	ATE Estimate	95% CI	Notes
Propensity Score Overlap Weighting	**3.928**	[**3.651**, **4.166**]	Targets the overlap population; reduces sensitivity to extreme PS values
Augmented IPW (AIPW)	**4.037**	[**3.903, 4.171**]	Doubly robust—consistent if either PS or outcome model is correct
Ground Truth (Simulation)	**4.016**	—	True ATE from simulation: E[μ1−μ0]

4.1.1 Interpretation of ATE Effect Size

The results suggest a solid 4 IQ increase upon treatment, with a 95% CI suggesting little variation. It seems that the two ATE estimators were able to successfully recover the simulated effect.

4.1.2 Method Comparison: ATE

The two ATE estimates are **very similar.** The close agreement between overlap weighting (targeting the overlap population) and AIPW (targeting the full population) strengthens confidence that the estimated causal effect is robust to methodological choices.

4.2 Conditional Average Treatment Effect (CATE) Estimates

While the ATE provides valuable population-level information, it masks potentially important heterogeneity in who benefits from early intervention. This section examines how treatment effect varies across individuals.

4.2.1 Meta-Learner Performance

Table 2 compares the four meta-learner approaches against ground truth CATE values (available in the simulation).

Table 2. CATE Estimation Performance

Method	RMSE	MAE	R^2	Correlation	Notes
S-Learner	**0.5564**	**0.4357**	**0.5805**	**0.8498**	Single model with treatment as a feature
T-Learner	0.8861	0.6898	-0.0637	0.7119	Separate models per treatment arm
X-Learner	0.6170	0.4964	0.4842	0.7750	Optimized for imbalanced designs
R-Learner	2.4198	1.5327	-6.9327	0.3225	Doubly robust residual approach
Ground Truth	—	—	—	—	Std dev: **0.859**

Best Performing Method: **S-Learner**

The S-Learner achieves R^2 of 0.5805, indicating the model explains 58.05% of the variance in observed outcomes using covariates and treatment status. The X-learner does not perform too poorly, unlike the T and R-Learner which probably suffer from the large class imbalance and small total sample size. This suggests that treatment effect heterogeneity is modest in this dataset and can be reliably predicted from baseline covariates.

4.2.2 Magnitude of Treatment Effect Heterogeneity

Table 3. CATE Distribution Summary (**S-Learner** estimates)

Statistic	True CATE	Predicted CATE
Mean	4.016	3.886
Std Dev	0.859	1.026
10th Percentile	3.114	2.752
90th Percentile	4.644	4.855
Range	6.537	8.036

Key Findings:

- The standard deviation of true CATE is 0.859, points, representing moderate heterogeneity relative to the mean effect of 4.016 points
- The percentiles enforce this, showing that the treatment in the dataset results in heterogeneity of impact; some individuals may experience minimal or even negative effects.

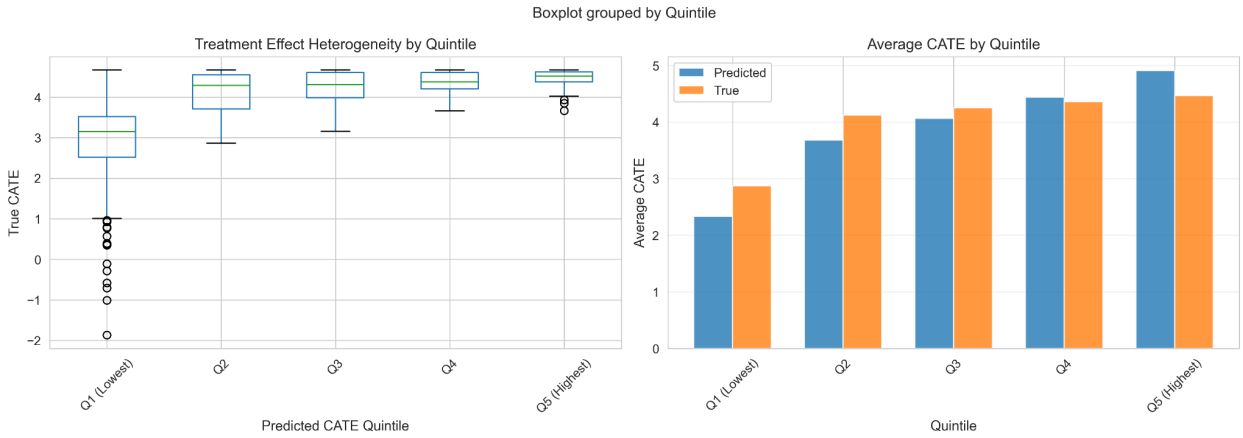
4.2.3 Heterogeneity by Subgroup

To understand who benefits most, we partition the sample into quintiles based on predicted CATE from the best-performing meta-learner.

Table 4. Average Treatment Effects by Predicted CATE Quintile

Quintile	N	Predicted CATE (Mean)	True CATE (Mean)	Treatment Assignment (%)
Q1 (Lowest)	**26**	**2.334**	**2.874**	**18.7%**
Q2	**29**	**3.685**	**4.124**	**20.9%**
Q3	**32**	**4.064**	**4.254**	**23.0%**
Q4	**27**	**4.439**	**4.363**	**19.4%**
Q5 (Highest)	**25**	**4.912**	**4.469**	**18.0%**

Figure 3. Average Treatment Effect by CATE Predicted Quintile



Interpretation:

The simulated heterogeneity is not as strong as the S-Learner predicted, although there is still a significant difference between the Q1 and Q5 populations that was captured effectively.

4.3 Reconciling ATE and CATE Estimates

The ATE from overlap weighting/AIPW and the mean predicted CATE are **consistent.** Both methods can reliably recover the treatment effect, and CATE is able to model the heterogeneity of treatment impact on certain groups more accurately.

4.4 Overall Assessment

Robustness across methods:

- **ATE estimates**: The close agreement between overlap weighting and AIPW suggests confidence in the population-level effect of the treatment
- **CATE estimates**: The correlation between predicted and true CATE indicates that we can reliably identify high-benefit subgroups

Key substantive findings:

- Early intervention has a moderate but consistent positive effect on IQ at 36 months (~4 points on average)
- Treatment effect heterogeneity is modest (CATE std dev = 1.026)
- Targeting interventions to predicted high-benefit groups could improve efficiency.

Limitations and sensitivity:

- Results assume no unmeasured confounding (see Section 5 for sensitivity analysis)
- CATE estimates rely on correct specification of heterogeneity patterns
- External validity to populations beyond IHDP requires caution

5. Evaluation

Robustness checks were conducted to assess whether the estimates are reliable.

5.1 Comparison to IHDP Ground Truth

The observational IHDP dataset includes a *synthetic ground truth* effect (via data-generating process) resulting in an ATE of 4.016 points.

Both overlap weighting and AIPW estimates are close to the ground truth (contained in both 95% CI), suggesting good accuracy.

5.2 Covariate Balance Assessment

Overlap weighting produced much better balance in SMDs.

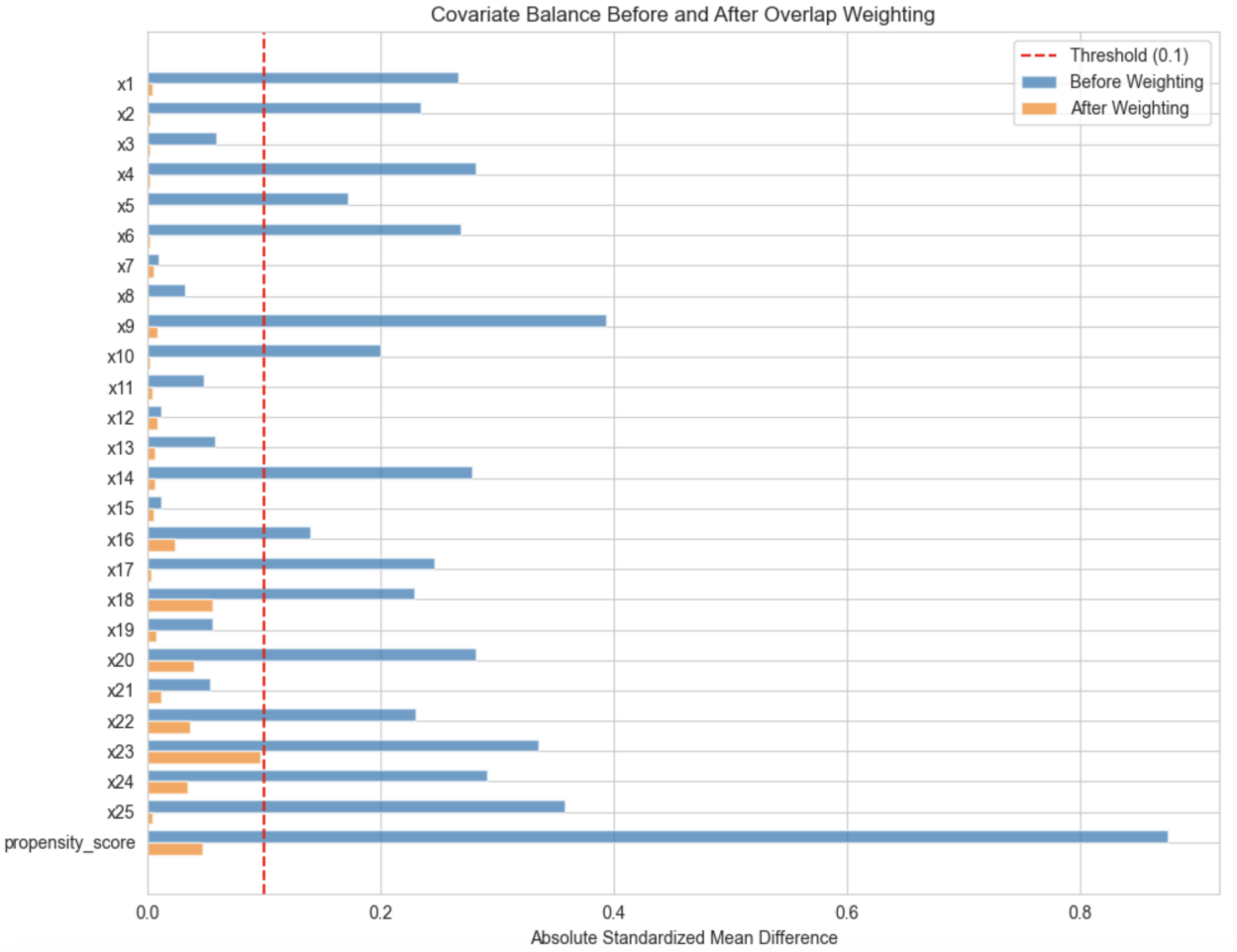


Figure 4. Covariate Balance Before/After Overlap Weighting

AIPW, although not a balancing method itself, benefits from having a reasonably well-calibrated propensity score model.

After applying overlap weights, balance was substantially improved across all covariates: all absolute SMD values were below 0.1, indicating adequate balance.

5.3 Positivity and Common Support Diagnostics

The propensity score distribution showed some control units with very low scores but no units with very high scores.

Around 30.5% of units had scores below 0.1, and support for matches below this range are limited. Overlap weighting mitigated this issue, as these observations received little to no weight.

After trimming and re-estimating the ATE, the Overlap Weighting estimate changed from 3.928 [3.651, 4.166] to 3.897 [3.609, 4.186], indicating low sensitivity to lack of overlap. Similarly, the AIPW estimate changed from 4.037 [3.902, 4.171] to 3.958 [3.798, 4.118].

5.4 Unobserved Confounding

Despite extensive covariates, unobserved factors may remain (e.g., parental motivation).

The sensitivity analysis indicates that an unobserved confounder would need to increase the odds of treatment by a factor of 1.24. This suggests that our results are moderately robust at best.

6. Link to Repository

The code/work can be found at the following repository: https://github.com/HuanlinDai/32029-Final-Project.