Understanding Causal Inference with Statistical and Machine Learning Models

Sunny Raja Prasad

Roll Number: 218171078

MS PROJECT 1(MTH697A)

Under the Supervision of

Professor Sharmishtha Mitra

February 2025

Acknowledgment

I would like to express my deepest gratitude to professor **Professor Sharmishtha Mitra** of the Mathematics and Statistics department for her unwavering guidance, encouragement, and invaluable mentorship throughout the completion of this project.

Professor
Mathematics and Statistics

Declaration

I hereby declare that the work presented in the project report entitled ”**Understanding Causal Inference with Statistical and Machine Learning Models**” contains my own ideas in my own words. At places, where ideas and words are borrowed from other sources, proper references, as applicable, have been cited. To the best of our knowledge this work does not emanate from or resemble other work created by person(s) other than mentioned herein.

Sunny Raja Prasad
Roll no- 218171078
BS-MS MTH

Contents

Abstract—**This study examines the impact of specialized childcare on cognitive test scores in premature infants using the Infant Health and Development Program (IHDP) dataset. We employ advanced causal inference techniques, including Double Machine Learning, Counterfactual Regression and Inverse Probability Weighting, to estimate treatment effects. Our findings suggest a positive impact of specialized childcare on cognitive outcomes, with implications for early intervention policies. The study reveals significant selection bias and limited overlap between treated and control groups, highlighting the importance of robust causal inference techniques in observational studies. Our analysis demonstrates the potential benefits of early interventions for vulnerable populations and underscores the need for careful methodological considerations in estimating causal effects from observational data.**

I Introduction

I-A Background and Motivation

Early childhood interventions, particularly for premature infants, have gained significant attention in recent years due to their potential long-term impacts on cognitive development. The Infant Health and Development Program (IHDP) provides a unique opportunity to study these effects in a rigorous manner. This research aims to contribute to the growing body of knowledge on early childhood interventions and their causal effects on cognitive outcomes.

I-B Research Question and Objectives

The primary research question addressed in this study is: ”Does specialized childcare improve cognitive test scores in premature infants?” To answer this question, we set the following objectives:

To estimate the causal effect of specialized childcare on cognitive test scores using advanced statistical and machine learning methods.

To compare and contrast different causal inference techniques in the context of the IHDP dataset.

To analyze potential heterogeneity in treatment effects across different subgroups of premature infants.

I-C Significance of the Study

Understanding the causal impact of specialized childcare on cognitive outcomes in premature infants is crucial for several reasons:

Policy Implications: Results from this study can inform early childhood education policies and resource allocation decisions.

Clinical Practice: Findings may influence recommendations for care and intervention strategies for premature infants.

Methodological Contributions: By comparing different causal inference techniques, this study contributes to the ongoing discussion on best practices in causal analysis of observational data.

I-D Fundamental Problem of Causal Inference

The fundamental problem of causal inference is that for any individual, we can only observe one potential outcome at a time. That is, we cannot simultaneously observe what would have happened both with and without the treatment for the same individual. This creates a missing data problem, making causal inference inherently challenging.

Mathematical Formulation: Let:

- $Y(1)$: Potential outcome if the individual receives the treatment.
- $Y(0)$: Potential outcome if the individual does not receive the treatment.
- T: Treatment assignment ($T = 1$ if treated, $T = 0$ if not).

The observed outcome is:

$$Y = TY(1) + (1 - T)Y(0)$$

The causal effect for an individual is:

$$Y(1) - Y(0)$$

However, we can only observe one of these values—either $Y(1)$ (if treated) or $Y(0)$ (if not treated). The other outcome remains counterfactual (i.e., we can never directly observe it). This missing counterfactual makes it impossible to compute individual causal effects directly.

1) Implications of the Fundamental Problem

1) **Causal Effects Can Only Be Estimated, Not Observed Directly:** Since we never observe both $Y(1)$ and $Y(0)$ for the same person, we need statistical methods (e.g., randomized controlled trials, matching, instrumental variables, etc.) to estimate causal effects.
2) **The Need for Assumptions:** To estimate causal effects, we often rely on assumptions such as:
   - **Randomization:** Ensures treatment assignment is independent of potential outcomes.
   - **Ignorability** ($Y(1), Y(0) \perp T|X$)**:** Given observed covariates X, treatment assignment is as good as random.
   - **Stable Unit Treatment Value Assumption (SUTVA):** No interference between units (one person's treatment does not affect another's outcome).
3) **Average Treatment Effects (ATE) as a Solution:** Since we cannot measure individual causal effects, we estimate the average treatment effect (ATE):

   $$ATE = E[Y(1) - Y(0)]$$

   This measures the expected difference in outcomes if we could treat and not treat the same individuals in two parallel universes.
4) **How RCTs Solve This Problem:** Randomized Controlled Trials (RCTs) help mitigate this problem by randomly assigning treatment, ensuring that the treated and control groups are statistically identical on average. This allows us to use the control group's observed outcomes as an estimate of the counterfactual outcomes for the treated group:

   $$E[Y(0)|T = 1] \approx E[Y(0)|T = 0]$$

   Thus, the observed difference in means between the treatment and control groups gives an unbiased estimate of the causal effect:

   $$E[Y(1)|T = 1] - E[Y(0)|T = 0] \approx ATE$$

I-E Key Definitions

To better understand the methodology and analysis presented in this report, we define some key concepts:

- **Treated Group:** The treated group consists of individuals who received the intervention or treatment being studied (e.g., specialized childcare in this case). Their outcomes are analyzed to assess the impact of the treatment.
- **Control Group:** The control group consists of individuals who did not receive the intervention or treatment. This group serves as a baseline for comparison to determine the causal effect of the treatment.

II Key Points About the IHDP Dataset

II-A Study Focus

The IHDP study aimed to investigate the effect of home visits on the cognitive development of premature infants. This research is crucial for understanding how early interventions can influence long-term cognitive outcomes in vulnerable populations.

II-B Data Access

Researchers can access the IHDP data through the Inter-university Consortium for Political and Social Research (ICPSR), a prominent repository for social science research data. The ICPSR provides a centralized platform for accessing and sharing datasets across various fields, facilitating collaborative research and analysis.

II-C Usage in Research

The IHDP dataset is widely used in causal inference research, particularly for benchmarking algorithms designed to estimate treatment effects. Its use in evaluating the performance of different causal inference methods, such as logistic regression with inverse probability weighting, double machine learning with random forests, and deep neural network-based counterfactual regression, makes it a valuable resource for researchers in this field.

III Causal Inference

III-A Overview

Causal inference is a statistical approach that aims to identify and quantify cause-and-effect relationships between variables. It goes beyond mere correlation to establish whether changes in one variable directly cause changes in another. This field is crucial in various disciplines, including economics, epidemiology, and social sciences, where understanding the true impact of interventions or treatments is essential.

III-B Correlation vs. Causation

The adage ”correlation does not imply causation” underscores a critical principle in statistical analysis: observed associations between variables do not inherently indicate causal relationships. This occurs due to:

- **Confounding Variables:** A third variable Z may influence both X and Y, creating spurious correlations. *Example:* Ice cream sales and drowning incidents correlate because both increase with temperature (Z), not because ice cream causes drowning.
- **Reverse Causation:** The direction of causality may be inverted. *Example:* Higher grades correlate with study hours, but innate ability (Z) may drive both.
- **Spurious Correlation:** Random chance can create meaningless associations. *Example:* Nicolas Cage films correlate with pool drownings, but no causal link exists.

These limitations motivate the need for causal inference methods like those employed in this study.

III-C Unobserved Confounding

Definition: Unobserved confounding occurs when hidden variables U influence both treatment T and outcome Y, biasing causal effect estimates. Mathematically, this violates ignorability:

$$Y(1), Y(0) \not\perp T \mid X$$

Examples:

- **Healthcare:** Unmeasured lifestyle factors may affect both drug adherence (T) and recovery (Y).
- **Economics:** Innate ability (U) may influence education (T) and earnings (Y).

Consequences:

- Biased ATE estimates due to violation of ignorability.
- Loss of identifiability (true causal effects cannot be estimated).
- Flawed policy decisions if unaddressed.

III-D Association vs. Causation

The distinction between association and causation is central to causal inference. While statistical methods can identify correlations, establishing causation requires overcoming three key challenges:

1) Definitions

- **Association (Correlation):** Two variables X and Y are associated if they co-occur or change together (e.g., measured via correlation coefficients).
- **Causation:** X causes Y if intervening on X systematically changes Y, holding all else constant (a counterfactual relationship).

2) Why Association is not Causation ?

1) **Confounding (Third Variable Problem):** A variable Z influences both X and Y. *Example:* Shoe size correlates with test scores because age (Z) drives both.
2) **Reverse Causation:** The assumed cause-effect direction is inverted. *Example:* Stress (Z) causes both coffee consumption (X) and poor sleep (Y).
3) **Spurious Correlation:** Random chance creates meaningless associations. *Example:* Cheese consumption correlates with bedsheet entanglement deaths.
4) **Selection Bias:** Non-random sampling creates illusory links. *Example:* Gym-goers appear healthier because healthier people choose to exercise.

These issues necessitate causal inference methods like IPW and DML to disentangle true causal effects from mere associations.

III-E Identifiability in Causal Inference

Identifiability refers to the ability to express causal quantities (e.g., Average Treatment Effect) in terms of observable data. A causal effect is identifiable if it can be uniquely computed from the observed data distribution under key assumptions.

Mathematical Formulation: The causal effect is defined as:

$$\tau = E[Y(1)] - E[Y(0)]$$

However, since only one of $Y(1)$ or $Y(0)$ is observed per individual, we rely on assumptions to estimate τ using observed data.

1) Identifiability Conditions

For causal effects to be identifiable, the following assumptions must hold:

1) **Ignorability (Unconfoundedness):**

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

Treatment assignment T is independent of potential outcomes given covariates X (no unmeasured confounders).

2) **Positivity (Overlap):**

$$0 < e(X) < 1 \quad \forall X$$

Every individual has a non-zero probability of receiving both treatment and control, where $e(X) = P(T = 1 \mid X)$ is the propensity score.

3) **Consistency:**

$$Y = TY(1) + (1 - T)Y(0)$$

Observed outcomes align with potential outcomes under the received treatment.

These assumptions enable estimation of causal effects from observational data, as discussed in our methodological approaches.

III-F Relevance to Our Problem

In the context of the Infant Health and Development Program (IHDP) dataset, causal inference is highly relevant for several reasons:

- **Treatment Effect Estimation**: We aim to determine the causal effect of specialized childcare on cognitive test scores in premature infants. This requires isolating the impact of the intervention from other confounding factors.
- **Addressing Selection Bias**: The IHDP dataset is observational, meaning that treatment assignment (specialized childcare) was not randomized. Causal inference methods help account for potential selection bias in treatment assignment.

- **Policy Implications**: Understanding the causal impact of early interventions on cognitive outcomes can inform policy decisions regarding childcare and early education programs for vulnerable populations.

III-G Literature Survey

The IHDP dataset has been widely used in causal inference research, particularly for benchmarking different estimation methods. Key studies include:

- **Hill (2011)**: Introduced Bayesian Additive Regression Trees (BART) for causal inference using the IHDP dataset, demonstrating its effectiveness in estimating treatment effects[1].
- **Shalit et al. (2017)**: Proposed Counterfactual Regression (CFR) using deep learning techniques, evaluating its performance on the IHDP dataset and showing improved estimation of individual treatment effects[2].
- **Künzel et al. (2019)**: Introduced Meta-learners for estimating heterogeneous treatment effects, using the IHDP dataset as one of their benchmark datasets[4].
- **Nie and Wager (2020)**: Developed Quasi-Oracle estimation methods for heterogeneous treatment effects, demonstrating their approach using the IHDP dataset among others[5].
- **Shi et al. (2019)**: Proposed adapting Generative Adversarial Networks (GANs) for causal inference, using the IHDP dataset to evaluate their method's performance in estimating average treatment effects[6].

IV Dataset Description

IV-A Overview

The Infant Health and Development Program (IHDP) dataset is based on a study conducted in the United States. The original IHDP study was designed as an early childhood intervention program aimed at improving the health and cognitive development of low-birth-weight, premature infants in the U.S.

The dataset used in causal inference research (like the one from the CEVAE repository) is a semi-synthetic version of the original IHDP study, where the outcomes are simulated based on a structural causal model. However, the pre-treatment covariates and study design originate from the real U.S.-based IHDP study. The IHDP dataset contains 747 observations from a non-randomized study, where pre-treatment covariates influence treatment assignment (specialized childcare). It includes 25 pre-treatment covariates, a binary treatment indicator, and an outcome variable (cognitive test scores). Due to the presence of confounding, causal inference techniques are required to estimate the treatment effect accurately.

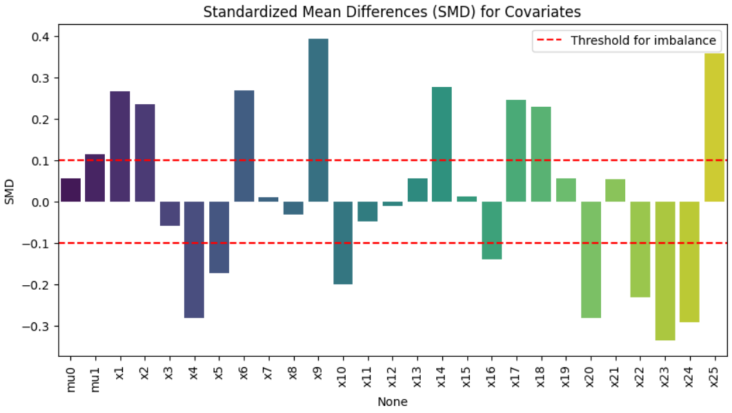

Fig. 1: Standardized Mean Differences (SMD) for covariates. Dashed line indicates acceptable imbalance threshold. Several covariates (particularly x1) show problematic imbalance.

- **Dashed Red Lines:** Represent the acceptable imbalance threshold (±0.1). Covariates with SMD values beyond these thresholds indicate imbalance between treatment and control groups.
- **Bars:** Each bar corresponds to a covariate (X1 to X25), showing its degree of imbalance.

- **Problematic Covariates:** Variables with SMD exceeding the threshold (particularly X1) indicate that treatment and control groups differ significantly on these characteristics.
- **Formula for SMD:** For a covariate X, the SMD is calculated as:

$$SMD = \frac{\bar{X}_T - \bar{X}_C}{s}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the means in treatment and control groups, and s is the pooled standard deviation:

$$s = \sqrt{\frac{s_T^2 + s_C^2}{2}}$$

- **Interpretation:**
  - $|SMD| < 0.1 \rightarrow$ Covariate is well balanced.
  - $0.1 \leq |SMD| < 0.25 \rightarrow$ Moderate imbalance.
  - $|SMD| \geq 0.25 \rightarrow$ Strong imbalance (indicates confounding risk).
- **Implications:** This imbalance necessitates adjustment methods like propensity score matching, inverse probability weighting (IPW), or covariate adjustment for proper causal inference.

IV-B Randomized Controlled Trials (RCTs): Key Theoretical Properties

Randomized Controlled Trials (RCTs) are considered the gold standard for evaluating causal relationships between interventions and outcomes. Below, we outline their key theoretical properties with mathematical justification:

1) **Elimination of Confounding:** Confounding occurs when both the treatment and outcome are influenced by a third variable. In an observational study, confounding is problematic because treatment assignment is not independent of pre-treatment covariates. However, in an RCT, randomization ensures that treatment assignment (T) is independent of confounders (X):

$$T \perp X$$

This eliminates systematic differences between treated and control groups due to confounders.

2) **Ensuring Exchangeability:** Exchangeability means that the distribution of potential outcomes is the same across treated and control groups, except for the treatment itself:

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

where $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively. This ensures valid causal inference by allowing the observed outcomes in the control group to represent counterfactual outcomes for the treated group.

3) **Enables Estimation of the Average Treatment Effect (ATE) Without Bias:** The Average Treatment Effect (ATE) is defined as:

$$ATE = E[Y(1) - Y(0)]$$

In an RCT, randomization ensures that:

$$E[Y(1)|T = 1] - E[Y(0)|T = 0] = E[Y(1) - Y(0)]$$

This means that the difference in observed means between treated and control groups is an unbiased estimator of the true causal effect.

4) **Eliminates Reverse Causation:** Reverse causation occurs when the outcome affects treatment assignment. In observational studies, this is a major issue (e.g., sicker individuals might be more likely to receive a treatment). However, in an RCT:

$$T \text{ is assigned independently of } Y(1), Y(0)$$

ensuring that treatment assignment precedes outcome observation, eliminating reverse causation.

5) **Eliminates Selection Bias:** Selection bias arises when individuals self-select into treatment based on characteristics related to the outcome. In an RCT, randomization ensures:

$$P(T = 1|X) = P(T = 0|X) = 0.5$$

for all covariates X, meaning that treatment assignment is independent of individual choices.

6) **Guarantees Unbiased Estimators (Law of Large Numbers):** By the Law of Large Numbers (LLN), as the sample size (n) increases, the sample mean approaches the population mean:

$$\hat{E}[Y|T=1] \to E[Y(1)], \tag{1}$$

$$\hat{E}[Y|T=0] \to E[Y(0)]. \tag{2}$$

ensuring that treatment effect estimates converge to their true values as n grows.

IV-C Key Variables

1) Outcome Variables

- Y_{factual} **(Observed Outcome):** This represents the observed cognitive test score for each individual given their actual treatment status.
  - If a child was in the treatment group, $Y_{\text{factual}} = Y(1)$.
  - If a child was in the control group, $Y_{\text{factual}} = Y(0)$.

  This is the only outcome directly observed in the dataset.

- Y_{cfactual} **(Counterfactual Outcome):** This represents the unobserved (counterfactual) cognitive test score—what the outcome would have been had the individual received the opposite treatment.
  - If a child was in the treatment group, $Y_{\text{cfactual}} = Y(0)$.
  - If a child was in the control group, $Y_{\text{cfactual}} = Y(1)$.

  This is never observed in real-world data, but in the IHDP dataset (a semi-synthetic dataset), counterfactuals are available through simulation.

2) Treatment Variable

- **Treatment** (T)**:** A binary indicator for whether a child received specialized childcare (treatment group) or not (control group).
  - $T = 1$: The child received treatment.
  - $T = 0$: The child was in the control group.

  This is the key variable for estimating causal effects.

3) Potential Outcomes (True Treatment Effects)

- μ_0 (Expected Outcome Without Treatment): Represents the expected cognitive test score if the child had not received specialized childcare.
  - For individuals where $T = 0$, this corresponds to their observed Y_{factual}.
  - For treated individuals ($T = 1$), this represents the counterfactual outcome.
- μ_1 (Expected Outcome With Treatment): Represents the expected cognitive test score if the child had received specialized childcare.
  - For individuals where $T = 1$, this corresponds to their observed Y_{factual}.
  - For untreated individuals ($T = 0$), this represents the counterfactual outcome.

4) Why These Variables Are Important?

These variables are critical for causal effect estimation:

1) The Individual Treatment Effect (ITE) is given by:

$$ITE = \mu_1 - \mu_0$$

2) The Average Treatment Effect (ATE) is given by:

$$ATE = E[\mu_1 - \mu_0]$$

These measures help estimate how much specialized childcare improves cognitive test scores on average and for specific individuals.

IV-D Covariate Explanation

The covariates can be categorized into several key groups, each relevant to the outcome in distinct ways:

1) *Child Characteristics*

- **Birth Weight** (continuous): Important for assessing health risks and developmental challenges. Low birth weight is associated with higher risks of cognitive and physical disabilities.
- **Health Indicators**:
  - X3: Head circumference (continuous) - Reflects brain development.
  - X4: Length at birth (continuous) - Indicates overall growth.
  - X5: Gestational age in weeks (continuous) - Affects maturity and health at birth.
  - X6: Neonatal health index (continuous) - Summarizes health status at birth.

2) *Maternal Characteristics*

- **Demographics**:
  - X7: Mother's age (continuous) - Influences parenting style and resource availability.
  - X8: Mother's race (categorical) - May reflect socioeconomic disparities.
  - X9: Mother's marital status (binary) - Affects family stability.
- **Socioeconomic Status**:
  - X10: Mother's educational attainment (ordinal) - Impacts access to resources and parenting quality.
  - X11: Mother's employment status (binary) - Influences financial stability and childcare options.
  - X12: Mother's work hours per week (continuous) - Affects childcare availability.
  - X13: Family income (continuous) - Determines access to healthcare and educational resources.

3) *Environmental Factors*

- **Household Characteristics**:
  - X14: Number of children in household (discrete) - Impacts resource allocation.
  - X15: Number of adults in household (discrete) - Affects support systems.
  - X16: Housing density (continuous) - Reflects living conditions.
- **Support Systems**:
  - X17: Social support index (continuous) - Influences stress levels and parenting quality.
  - X18: Father's presence (binary) - Contributes to family stability.
  - X19: Grandmother's presence (binary) - Provides additional support.

4) *Study-Specific Variables*

- **Site Characteristics**:
  - X20-X25: Site indicators (binary) - Represent different study locations.

IV-E Data Quality and Limitations

The IHDP dataset is generally well-structured, but several limitations and potential biases should be considered, as evidenced by the following analyses:

- **Selection Bias**: Significant covariate imbalance is evident in:

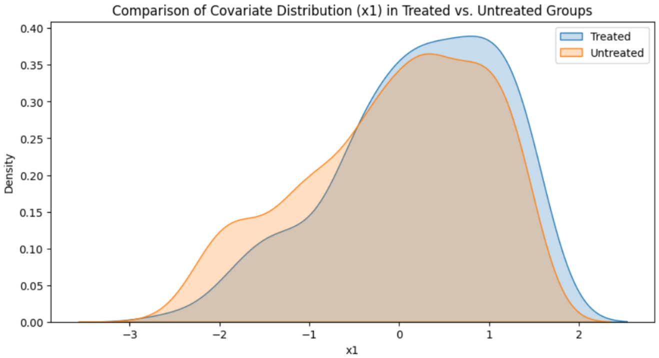

Fig. 2: Distribution of covariate x1 in treated vs. untreated groups, showing systematic differences that violate the randomization assumption.

- **Counterfactual Uncertainty**: Significant divergence between observed and estimated outcomes:

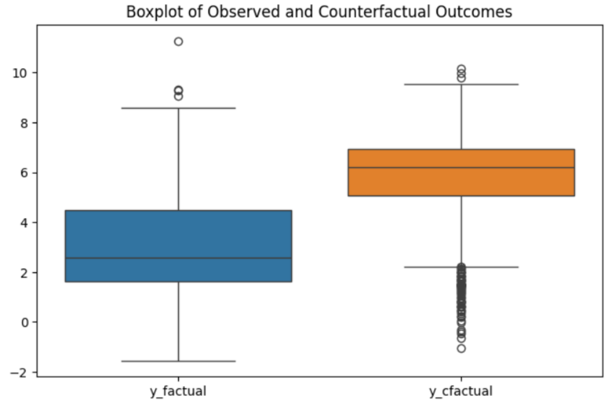

Fig. 3: Comparison of factual vs. counterfactual outcomes. The non-overlapping distributions suggest model sensitivity to unobserved confounding.

- **Missing Values**: There is no missing values in the given dataset.
- **Generalizability**: The combined evidence from:
  - Covariate imbalance (Figures 2, 1)
  - Counterfactual divergence (Figure 3)

  suggests limited external validity for populations with different characteristics.

These limitations highlight the need for:

- Propensity score weighting to address imbalance
- Sensitivity analyses for counterfactual estimates
- Explicit documentation of missing data mechanisms

IV-F Causal Graph

The causal graph was constructed based on the following assumptions and steps:

- **Assumptions**: We assumed that all 25 covariates are potential confounders and that there are no unmeasured confounders. The graph includes directed edges from these covariates to both the treatment (specialized childcare) and the outcome (cognitive test scores).
- **Construction**: The graph was built by identifying all backdoor paths from treatment to outcome and ensuring that these paths are blocked by conditioning on the set of covariates. This approach allows for the estimation of the causal effect of specialized childcare on cognitive outcomes.

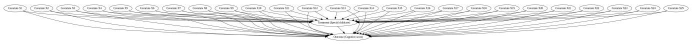

Fig. 4: Causal Graph showing relationships between Covariates (X1-X25), Treatment, and Outcome

The causal graph illustrates:

- **Structure**:
  - 25 covariates (X1-X25) as potential confounders
  - Treatment node representing specialized childcare
  - Outcome node representing cognitive scores
- **Relationships**:
  - Direct edges from covariates to both treatment and outcome
  - Direct causal effect from treatment to outcome
  - No direct connections between covariates (assuming conditional independence)

V Methodology

V-A Introduction to Average Treatment Effect (ATE)

The **Average Treatment Effect (ATE)** is a central concept in causal inference. It quantifies the expected difference in outcomes between the treatment and control groups across a population. Formally, the ATE is defined as:

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

where $Y(1)$ denotes the potential outcome if an individual receives the treatment and $Y(0)$ denotes the potential outcome if the individual does not receive the treatment.

Why Estimate the ATE?

In observational studies, such as those using the IHDP dataset to assess the impact of specialized childcare on cognitive test scores, treatment assignment is not randomized. This can lead to confounding bias if individuals receiving treatment differ systematically from those who do not. Estimating the ATE allows us to infer the causal effect of the treatment by comparing the average outcomes that would have been observed under both treatment and control conditions. Since we only observe one of these outcomes for each individual, various methods are employed to estimate the unobserved (counterfactual) outcome and thus the ATE.

Assumptions for Estimating ATE

When estimating the ATE using methods such as Inverse Probability Weighting (IPW), Double Machine Learning (DML), or Counterfactual Regression (CFR), we rely on several key assumptions:

1) **Ignorability (Unconfoundedness):**

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

This assumption states that, given a set of observed covariates X, the treatment assignment T is independent of the potential outcomes. In other words, there are no unobserved confounders that affect both the treatment and the outcome.

2) **Positivity (Overlap):**

$$0 < P(T = 1 \mid X) < 1 \quad \forall X$$

This ensures that every individual has a nonzero probability of receiving either treatment or control, making it possible to compare outcomes across groups.

3) **Stable Unit Treatment Value Assumption (SUTVA):**

$$Y_i = Y_i(T_i)$$

This assumption requires that the treatment assigned to one individual does not affect the outcomes of another, and that each treatment is well-defined.

ATE Estimation Methods

Different methods apply these assumptions to estimate the ATE:

1. Logistic Regression with Inverse Probability Weighting (IPW):

- **Estimation Process:**

1) **Propensity Score Estimation:** Estimate $p(t = 1 \mid x)$ using a logistic regression model.

$$p(t = 1 \mid x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$

2) **Stabilized Weights Calculation:** Compute weights as

$$w_i = \frac{T_i}{2u} + \frac{(1 - T_i)}{2(1 - u)},$$

where $u = \frac{1}{N} \sum_{i=1}^{N} T_i$ is the global treatment proportion.

3) **ATE Calculation:** Use the weighted outcomes to estimate the ATE:

$$\widehat{\text{ATE}}_{IPW} = \frac{\sum_i T_i Y_i}{\sum_i T_i p(t_i = 1 \mid x_i)} - \frac{\sum_i (1 - T_i) Y_i}{\sum_i (1 - T_i)(1 - p(t_i = 1 \mid x_i))}. \tag{3}$$

2. Double Machine Learning (DML):

- **Estimation Process:**

1) **Outcome Modeling:** Use a flexible model (e.g., RandomForestRegressor) to predict outcomes $\hat{Y}(X)$.

2) **Treatment Modeling:** Use a model (e.g., RandomForestClassifier) to estimate propen-

sity scores $\hat{P}(T = 1 \mid X)$.

3) **Residualization:** Compute residuals to remove the effects of X:

$$\tilde{Y}_i = Y_i - \hat{Y}(X_i), \quad \tilde{T}_i = T_i - \hat{P}(T = 1 \mid X_i).$$

4) **ATE Calculation:** Estimate the ATE as the ratio of the covariance between the residuals:

$$\widehat{\text{ATE}}_{DML} = \frac{E[\tilde{Y}_i \tilde{T}_i]}{E[\tilde{T}_i^2]}.$$

3. Deep Neural Network (Counterfactual Regression, CFR):

- **Representation Learning:** Map covariates X to a balanced latent space:

$$h(X) = \text{ReLU}\Big(W_2 \cdot \text{ReLU}(W_1 X + b_1) + b_2\Big),$$

where W_1, b_1 and W_2, b_2 are learned parameters.

- **Outcome Prediction:** Use two separate heads to predict potential outcomes:

$$\hat{Y}^1 = f_1(h(X)), \quad \hat{Y}^0 = f_0(h(X)).$$

- **ATE Calculation:** The Individual Treatment Effect (ITE) is:

$$\text{ITE}(X) = \hat{Y}^1 - \hat{Y}^0,$$

and the ATE is the average over the sample:

$$\widehat{\text{ATE}}_{CFR} = \frac{1}{N} \sum_{i=1}^{N} \text{ITE}(X_i).$$

Overall, these methods share the goal of isolating the causal effect of treatment on the outcome by adjusting for confounding. They all rely on the core assumptions of ignorability, positivity, and SUTVA to ensure that the estimated ATE is unbiased.

V-B Individual Treatment Effect (ITE) Estimation

V-C Definition

The Individual Treatment Effect (ITE) quantifies the causal effect of a treatment at an individual level. It is defined as:

$$ITE(X) = Y^1 - Y^0$$

where:

- Y^1 is the potential outcome if the individual receives the treatment.
- Y^0 is the potential outcome if the individual does not receive the treatment.

V-D Why is ITE Estimation Needed?

- **Heterogeneous Treatment Effects:** The treatment effect may vary across individuals, and ATE alone does not capture this variation.
- **Personalized Decision-Making:** Knowing ITE allows for targeted interventions where treatment is most beneficial.
- **Model Evaluation:** Many machine learning models estimate individual treatment effects, making ITE essential for assessing their performance.

V-E Assumptions for ITE Estimation

1) Inverse Probability Weighting (IPW)

- **Ignorability (Unconfoundedness):**

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

- **Positivity (Overlap):**

$$0 < P(T = 1 \mid X) < 1 \quad \forall X$$

- **Stable Unit Treatment Value Assumption (SUTVA):** Treatment assignment to one individual does not affect others.

2) Double Machine Learning (DML)

- Same assumptions as IPW.
- **Model Specification Assumption:** The outcome and treatment models must be correctly specified.

3) Counterfactual Regression (CFR)

- Same assumptions as IPW and DML.
- **Representation Learning Assumption:** There exists a latent representation of X that allows matching treated and control distributions.

V-F Why Different Methods Differ in ITE Estimation?

- **IPW:** Estimates ATE well but is not designed for individual-level effects.
- **DML:** Controls for confounders using machine learning, allowing better ITE estimation.
- **CFR:** Learns balanced representations to model both Y^1 and Y^0 effectively, making it most suitable for ITE estimation.

V-G Inverse Probability Weighting (IPW)

1) What is IPW and Why Use It?

Inverse Probability Weighting (IPW) is a causal inference method used to adjust for confounding in observational studies. It works by assigning weights to each observation based on their propensity scores, which represent the probability of receiving treatment given observed covariates. These weights create a pseudo-population where the distribution of covariates is balanced between treated and control groups, mimicking a randomized experiment.

Propensity Score: The propensity score is the probability of receiving treatment given a set of observed covariates. Formally, it is defined as:

$$e(X) = P(T = 1|X)$$

where:

- $e(X)$ is the propensity score
- T is the treatment assignment (1 = treated, 0 = control)
- X represents observed covariates

In causal inference, propensity scores help adjust for selection bias when estimating treatment effects.

Confounding: Confounding occurs when a third variable (a confounder) influences both the treatment (independent variable) and the outcome (dependent variable), causing a spurious association between them. In causal inference, confounding can bias the estimation of the true causal effect unless adjustments are made using methods like propensity score matching or regression.

Propensity score matching: Propensity Score Matching (PSM) is a statistical technique used to estimate the causal effect of a treatment by matching treated and control units with similar propensity scores. It helps balance observed covariates between the two groups, mimicking a randomized experiment.

Mathematical Proof of the Propensity Score Theorem(Based on Rosenbaum & Rubin (1983)):

Theorem

If treatment assignment T is strongly ignorable given covariates X, i.e.,

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

here Y is potential outcome, then it is also strongly ignorable given the propensity score $e(X)$, i.e.,

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

where the propensity score is defined as:

$$e(X) = P(T = 1 \mid X).$$

Proof

Step 1: Factorization of Joint Distribution

We can factorize the joint probability as follows:

$$\begin{aligned} P(Y(0), Y(1), T \mid X) = P(Y(0), Y(1) \mid X) \\ \times e(X)^T (1 - e(X))^{1-T} \end{aligned}$$

This factorization uses the assumption of strong ignorability given X, which allows us to write:

$$\begin{aligned} P(T \mid Y(0), Y(1), X) &= P(T \mid X) \\ &= e(X)^T (1 - e(X))^{1-T} \end{aligned}$$

Step 2: Marginalization Over X

Now, we marginalize over X to get the joint probability of $(Y(0), Y(1), T)$ given $e(X)$:

$$\begin{aligned} &P(Y(0), Y(1), T \mid e(X)) = \\ &\int P(Y(0), Y(1), T \mid X) P(X \mid e(X)) dX \end{aligned}$$

Substituting the equation from Step 1:

$$\begin{aligned} &P(Y(0), Y(1), T \mid e(X)) = \\ &\int P(Y(0), Y(1) \mid X) e(X)^T \\ &\quad \times (1 - e(X))^{1-T} P(X \mid e(X)) dX \end{aligned}$$

Step 3: Factoring out Treatment Terms

Since $e(X)^T (1 - e(X))^{1-T}$ is a function of $e(X)$ and T only, we can factor it out of the integral:

$$\begin{aligned} &P(Y(0), Y(1), T \mid e(X)) = \\ &e(X)^T (1 - e(X))^{1-T} \\ &\quad \times \int P(Y(0), Y(1) \mid X) P(X \mid e(X)) dX \end{aligned}$$

Step 4: Defining Conditional Distribution

Define:

$$\begin{aligned} &P(Y(0), Y(1) \mid e(X)) = \\ &\int P(Y(0), Y(1) \mid X) P(X \mid e(X)) dX \end{aligned}$$

This is the conditional distribution of potential outcomes given the propensity score.

Step 5: Final Factorization

With this definition, we can rewrite the joint distribution as:

$$\begin{aligned} &P(Y(0), Y(1), T \mid e(X)) = \\ &\quad P(Y(0), Y(1) \mid e(X)) P(T \mid e(X)) \end{aligned}$$

Where $P(T \mid e(X)) = e(X)^T (1 - e(X))^{1-T}$.

This factorization demonstrates that, conditional on the propensity score $e(X)$, the potential outcomes $(Y(0), Y(1))$ are independent of the treatment assignment T. This is exactly the definition of strong ignorability given $e(X)$:

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

Therefore, if treatment assignment is strongly ignorable given covariates X, then it is also strongly ignorable given just the propensity score $e(X)$.

Intuition Behind the Proof

1) **Start with ignorability:** The treatment assignment T is independent of potential outcomes given all covariates X.
2) **Express the joint distribution:** Use the law of total probability and the definition of the propensity score $e(X)$.
3) **Marginalize over X:** Show that conditioning on $e(X)$ preserves independence.
4) **Conclusion:** The propensity score contains all relevant information about treatment assignment, making it sufficient for causal inference.

2) Intuition for Propensity Scores

The propensity score $e(W) = P(T = 1 \mid W)$ serves as a sufficient statistic for covariate adjustment. Instead of conditioning on the full covariate vector W, we use $e(W)$ because:

1) **Balancing Property:** Conditioning on $e(W)$ balances covariates between treated and control groups:

$$W \perp T \mid e(W)$$

This reduces confounding by making groups comparable, even with high-dimensional W.

2) **Dimensionality Reduction:** Propensity scores collapse W into a single scalar $e(W)$, simplifying adjustment (e.g., weighting or matching).

3) **Sufficiency:** $e(W)$ retains all information about W relevant to treatment assignment. Adjusting for $e(W)$ is as effective as adjusting for W for removing bias.

4) **Overlap Enforcement:** Focuses analysis on units with $0 < e(W) < 1$, avoiding extrapolation where treated/control groups lack overlap.

Example: In a job training program study, W might include education, experience, and income. Instead of adjusting for all three, we use $e(W)$ (the probability of enrollment) to balance groups.

Key Takeaway: Propensity scores make causal inference tractable by summarizing confounding information into a single metric while preserving validity.

Residual Confounding: Residual confounding occurs when covariates remain imbalanced between treated and control groups even after applying adjustment methods like IPW. This imbalance can introduce bias into the estimated Average Treatment Effect (ATE). For example:

- Standardized Mean Difference (SMD) greater than 0.1 after IPW suggests that some covariates are not well balanced.
- The treated and control groups may still differ in ways unrelated to the treatment.
- The ATE estimate might reflect a mix of treatment effects and confounding effects rather than the pure causal effect.

Why Use IPW?

- **Address Confounding:** IPW helps control for confounding by balancing covariates between treated and control groups, ensuring unbiased estimation of treatment effects.
- **Handle Selection Bias:** By reweighting observations, IPW accounts for selection bias in treatment assignment.
- **Estimate Average Treatment Effect (ATE):** IPW facilitates estimation of the ATE by comparing weighted outcomes between treated and control groups.

Propensity Score Foundation: Inverse Probability Weighting (IPW) uses estimated propensity scores to reweight observations:

$$e(X) = P(T=1|X) = \frac{\exp(\beta_0 + \sum_{j=1}^{25} \beta_j X_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{25} \beta_j X_j)} \tag{4}$$

where $X_1, ..., X_{25}$ are the 25 pre-treatment covariates[2][5].

Weighting Strategy:

$$w_i = \begin{cases} \frac{1}{e(X_i)} & \text{if } T_i = 1 \\ \frac{1}{1-e(X_i)} & \text{if } T_i = 0 \end{cases} \tag{5}$$

Creates a pseudopopulation where treatment assignment becomes independent of covariates [6] [8].

Covariate Balance: IPW equalizes covariate distributions between groups through:

$$P(T=1|X,w) \approx P(T=0|X,w) \tag{6}$$

Achieving balance comparable to randomized trials when weights are correctly specified[2][5].

3) Assumptions for Causal Inference

To ensure valid causal effect estimation using IPW, the following assumptions must hold:

1) **Ignorability (Unconfoundedness):**

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

Treatment assignment T is independent of potential outcomes $Y(0)$, $Y(1)$ given covariates X. This assumes no unmeasured confounders.

2) **Positivity (Overlap):**

$$0 < e(X) < 1 \quad \forall X$$

Every individual has a non-zero probability of receiving both treatment and control, ensuring overlap in covariate distributions.

3) **Consistency:**

$$Y = TY(1) + (1 - T)Y(0)$$

Observed outcomes correspond to the potential outcomes under the received treatment (no misclassification).

4) **Correct Propensity Score Model Specification:**

$$\hat{e}(X) \approx P(T = 1 \mid X)$$

The estimated propensity score $\hat{e}(X)$ must accurately reflect the true treatment probability to avoid biased weighting.

ATE Estimation: Unbiased treatment effect calculation via:

$$\hat{\tau}_{IPW} = E\left[\frac{TY}{e(X)}\right] - E\left[\frac{(1-T)Y}{1-e(X)}\right] \tag{7}$$

where $\hat{\tau}_{IPW}$ represents the estimated causal effect of the treatment on the outcome using the IPW method. It corrects for confounding bias by weighting individuals based on their propensity scores $e(X)$, ensuring that treated and untreated groups are comparable. The treatment assignment variable T is binary, indicating whether an individual received the intervention:

- $T = 1$ if the infant received the early childhood intervention treatment (e.g., educational and medical support).
- $T = 0$ if the infant did not receive the intervention (control group).

The outcome variable Y used in the IPW formula refers to the observed **factual** outcome Y_{factual}, since we only observe one potential outcome per individual:

- If $T = 1$, then $Y_{\text{factual}} = Y(1)$, the outcome under treatment.
- If $T = 0$, then $Y_{\text{factual}} = Y(0)$, the outcome under control.

Since counterfactual outcomes Y_{cfactual} are unobserved, they are not used in the estimation process but can be leveraged for evaluation in semi-synthetic datasets like IHDP.

4) *Logistic Regression for Propensity Score Estimation*

The IPW method uses logistic regression to estimate propensity scores, which represent the probability of receiving treatment (specialized childcare) given the observed covariates. The mathematical formulation is:

$$P(T = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_{25} X_{25})}}$$

Where:

- T is the binary treatment indicator (1 for specialized childcare, 0 for control)
- $X_1, \ldots, X_{25}$ are the 25 pre-treatment covariates
- $\beta_0, \ldots, \beta_{25}$ are the regression coefficients to be estimated

5) *Mathematical Justification: MLE for Propensity Score Estimation*

We model the probability of receiving treatment as:

$$P(T_i = 1|X_i) = \pi_i = \frac{1}{1 + e^{-X_i^T \beta}}$$

where:

- $T_i \sim Bernoulli(\pi_i)$: The treatment assignment follows a Bernoulli distribution with probability π_i.
- X_i: The covariates for individual i.
- β: The vector of logistic regression coefficients to be estimated.

Since each T_i is a Bernoulli random variable, the likelihood function for the dataset (assuming independence across samples) is:

$$L(\beta) = \prod_{i=1}^{N} \pi_i^{T_i}(1-\pi_i)^{1-T_i}$$

Taking the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{N} [T_i \log \pi_i + (1-T_i)\log(1-\pi_i)]$$

Substituting $\pi_i = \frac{1}{1+e^{-X_i^T\beta}}$, we get:

$$\begin{aligned}\ell(\beta) = \sum_{i=1}^{N} \Bigg[& T_i \log \frac{1}{1+e^{-X_i^T\beta}} \\ & + (1-T_i)\log\left(1-\frac{1}{1+e^{-X_i^T\beta}}\right)\Bigg].\end{aligned} \tag{8}$$

This is the standard logistic regression likelihood function. To estimate β, we maximize this log-likelihood using optimization techniques such as gradient descent or Newton-Raphson.

Conclusion: - The treatment variable follows a Bernoulli distribution, and we assume independence of treatment assignments. - Maximum Likelihood Estimation (MLE) is used to estimate the coefficients (β) of the logistic regression model. - The estimated propensity scores (π_i) are then used for Inverse Probability Weighting (IPW) in causal inference.

V-H Implementation details:

- Maximum iterations: 1000
- Propensity score clipping: Stabilization range set to [0.1, 0.9]

The maximum number of iterations (1000) in our implementation refers to the optimization process:

1) The algorithm starts with initial coefficient guesses.
2) It iteratively adjusts these coefficients to maximize the likelihood of observing the data.
3) This process continues until either the model converges or the maximum number of iterations is reached.

Setting a maximum of 1000 iterations balances allowing sufficient time for convergence while ensuring the process terminates in a reasonable timeframe. The propensity score clipping (stabilization range [0.1, 0.9]) helps prevent extreme weights and improve numerical stability in the subsequent analysis.

V-I Double Machine Learning (DML)

1) What is DML?

Double Machine Learning (DML) is a causal inference method that combines machine learning models with econometric techniques to estimate treatment effects while controlling for confounding variables. It is particularly useful in high-dimensional settings where there are many covariates, allowing for flexible modeling of both outcomes and treatment assignments.

Key Features of DML:

- **Two-Stage Procedure:**
  - In the first stage, machine learning models predict the outcome and treatment probabilities (propensity scores) based on covariates.
  - In the second stage, residualized versions of the outcome and treatment are used to estimate the Average Treatment Effect (ATE).

- **Flexibility:** DML can handle nonlinear relationships between covariates and outcomes/treatment assignments.
- **Double Robustness:** The method provides consistent estimates even if one of the models (outcome or treatment) is misspecified.

2) What is Random Forest?

Random Forest is a machine learning algorithm that builds multiple decision trees during training and combines their outputs to make predictions. It is widely used for both regression and classification tasks due to its robustness and ability to handle complex data structures.

Key Features of Random Forest:

- **Ensemble Learning:** Combines predictions from multiple decision trees to improve accuracy and reduce overfitting.
- **Nonlinear Modeling:** Captures complex, nonlinear relationships between input features and target variables.
- **Feature Importance:** Provides insights into which covariates are most important for predicting the outcome or treatment assignment.
- **Robustness:** Handles missing data, outliers, and high-dimensional datasets effectively.

In this study, Random Forests are used as:

- **Outcome Model**: ‘RandomForestRegressor‘ predicts cognitive test scores based on covariates.
- **Treatment Model**: ‘RandomForestClassifier‘ predicts the probability of receiving specialized childcare based on covariates.

3) What Are Nonlinear Relationships?

A nonlinear relationship occurs when the relationship between two variables cannot be accurately described using a straight line. Instead, it may follow a curve or other complex pattern.

Examples of Nonlinear Relationships:

- Quadratic Relationship: $y = ax^2 + bx + c$
- Exponential Relationship: $y = ae^{bx}$
- Interaction Effects: When the effect of one variable on the outcome depends on the value of another variable.

In causal inference, nonlinear relationships often arise between covariates and outcomes or treatment assignments. For instance:

- The effect of income on cognitive test scores might increase at a diminishing rate (e.g., logarithmic relationship).
- The probability of receiving specialized childcare might depend on complex interactions between maternal education, household income, and geographic location.

Random Forests are particularly suited for capturing such nonlinear relationships because they partition the feature space into regions using decision trees, allowing them to model complex patterns without requiring explicit specification of the functional form.

4) Implementation Details

The DML framework uses machine learning models to estimate treatment effects while controlling for confounding variables. We chose Random Forests for both the outcome and treatment models due to their ability to handle high-dimensional data and capture nonlinear relationships without extensive parameter tuning [7]. Random Forests are robust against overfitting and can efficiently manage complex interactions among the 25 covariates in our dataset [**?**].

Why Use RandomForestRegressor for Outcome Model? The outcome model predicts cognitive test scores (a continuous variable). RandomForestRegressor was chosen because:

- **Nonlinear Modeling:** Captures complex relationships between covariates and cognitive test scores.
- **Feature Importance:** Provides insights into which covariates are most influential in predicting outcomes.
- **Robustness:** Handles high-dimensional data effectively and avoids overfitting.
- **Flexibility:** Does not require assumptions about functional relationships between variables.

Why Use RandomForestClassifier for Treatment Model? The treatment model predicts whether an individual received specialized childcare (binary variable). RandomForestClassifier was chosen because:

- **Binary Classification:** Suitable for predicting binary outcomes such as treatment assignment ($T = 0$ or $T = 1$).
- **Nonlinear Decision Boundaries:** Models complex interactions between covariates influencing treatment assignment.
- **Robustness:** Avoids overfitting and handles high-dimensional datasets effectively.
- **Probabilistic Outputs:** Provides probabilities ($P(T = 1)$) used as propensity scores in causal inference methods like IPW.

What Is GridSearchCV? GridSearchCV is a hyperparameter tuning method that systematically searches through a predefined set of hyperparameters to find the combination that yields the best performance. The process involves:

- Defining a grid of hyperparameters (e.g., number of trees, maximum depth).
- Performing cross-validation to evaluate model performance on unseen data.
- Selecting the combination of hyperparameters that achieves the best score across all validation folds.

GridSearchCV was used in this study to optimize hyperparameters such as: **1. Number of Trees ($n_{\textbf{estimators}} = 100$):**

- The model aggregates predictions from 100 decision trees to reduce variance and improve generalizability.
- Trade-off: Higher $n_{\text{estimators}}$ increases computational cost with diminishing returns. 100 trees balances robustness and efficiency.
- Justification: Empirical studies suggest convergence in performance around 100-200 trees for many datasets [1][2].

2. Random State ($random\_state = 42$):

- Ensures reproducibility by fixing the random seed for bootstrapping and feature selection.
- Without this, results would vary across runs due to inherent randomness in tree construction.

5) Backdoor Adjustment Framework

To estimate the causal effect of treatment T (specialized childcare) on outcome Y (cognitive scores), we use the backdoor adjustment criterion [3]. This blocks non-causal paths between T and Y by conditioning on confounders X:

$$P(Y \mid do(T)) = \sum_{X} P(Y \mid T, X)P(X)$$

Key Components:

- $Y = Y^{\text{factual}}$: Observed outcome from the IHDP dataset.
- $T \in \{0, 1\}$: Treatment indicator (1 = treated, 0 = control).
- X: Pre-treatment covariates (e.g., birth weight, maternal education).

6) Observational vs. Interventional Treatment

1. Observational Treatment Assignment (T): The variable T represents the treatment as observed in real-world data:

- $T = 1$: Individual received treatment (e.g., specialized childcare).
- $T = 0$: Individual did not receive treatment.

The observed relationship $P(Y \mid T)$ captures **correlation** but may be confounded by covariates X (e.g., socioeconomic status).

2. Interventional Treatment Assignment ($do(T)$): The $do(T)$ operator [3] represents an idealized experiment where T is set externally (e.g., randomized trial):

$$P(Y \mid do(T)) = \text{Causal effect of } T \text{ on } Y$$

This intervenes to remove confounding paths, isolating the direct effect of T on Y.

Key Difference:

$$\begin{aligned} P(Y \mid T) &\quad \text{(Correlation: Observed association)} \\ P(Y \mid do(T)) &\quad \text{(Causation: Unconfounded effect)} \end{aligned}$$

7) Model Architecture

We employ two Random Forest models to estimate causal effects:

- **Outcome Model (RandomForestRegressor):** Predicts the continuous outcome Y (cognitive test scores) using covariates X:

  $$\hat{Y}(X) = \text{RandomForestRegressor}(X)$$

  Captures nonlinear relationships between X and Y through ensemble decision trees.
- **Treatment Model (RandomForestClassifier):** Estimates propensity scores $P(T = 1|X)$ (probability of receiving specialized childcare):

  $$\hat{P}(T = 1|X) = \text{RandomForestClassifier}(X)$$

  Classifies treatment assignment $T \in \{0, 1\}$ while adjusting for confounders.

8) ATE Estimation via Residualization

The DML procedure involves two stages:

1) **Residualization:**
   - Outcome residual: $\tilde{Y}_i = Y_i - \hat{Y}(X_i)$
   - Treatment residual: $\tilde{T}_i = T_i - \hat{P}(T = 1|X_i)$

     Y is the observed outcome from y_{factual}, and the residuals remove the effect of covariates X from Y and T.
2) **ATE Calculation:**

   $$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tilde{Y}_i \cdot T_i}{\hat{P}(T = 1|X_i)} - \frac{\tilde{Y}_i \cdot (1 - T_i)}{1 - \hat{P}(T = 1|X_i)} \right)$$

   where ATE is average treatment effect by DML approach and this weighted average compares residualized outcomes between treated and control groups.

9) Intuition Behind DML

- **Outcome Model:** Isolates the part of Y unexplained by X, focusing on treatment-induced variation.
- **Treatment Model:** Adjusts for selection bias by modeling treatment likelihood given X.
- **Residualization:** Removes confounding by ”filtering out” covariate effects before estimating ATE.

V-J Counterfactual Regression (CFR)

1) Model Architecture

2) Definition and Key Idea

Counterfactual Regression (CFR) is a machine learning framework for estimating causal effects by learning balanced representations of treated ($T = 1$) and control ($T = 0$) groups. Unlike methods that rely solely on observational associations, CFR predicts both **factual** and **counterfactual** outcomes:

- **Factual Outcome (Y^F):** Observed outcome under the actual treatment (e.g., $Y^F = Y(1)$ if $T = 1$).
- **Counterfactual Outcome (Y^C):** Unobserved outcome under the opposite treatment (e.g., $Y^C = Y(0)$ if $T = 1$).

The goal is to estimate the **Individual Treatment Effect (ITE)** for each individual:

$$\text{ITE}_i = \hat{Y}_i(T = 1) - \hat{Y}_i(T = 0)$$

and the **Average Treatment Effect (ATE)** across the population:

$$\text{ATE} = \frac{1}{N} \sum_{i=1}^{N} \text{ITE}_i$$

3) Balanced Representation Learning

CFR learns a shared latent representation $h = f(X)$ that:

- Encodes covariate information X into a low-dimensional space.
- Balances treated and control distributions to mimic randomization.
- Enables accurate prediction of both $Y(1)$ and $Y(0)$ through separate outcome heads.

4) Advantages Over Traditional Methods

- **Heterogeneous Effects:** Captures variation in treatment effects across subgroups.
- **Confounder Adjustment:** Uses Maximum Mean Discrepancy (MMD) to align treated/control representations.
- **Nonlinear Relationships:** Leverages neural networks to model complex interactions.

5) Maximum Mean Discrepancy (MMD) Loss

Definition and Purpose: MMD Loss measures the discrepancy between the latent representations of treated ($T = 1$) and control ($T = 0$) groups. By minimizing MMD, CFR aligns their distributions in the representation space, reducing confounding bias.

Mathematical Formulation: Given treated ($Z_t = f(X_t)$) and control ($Z_c = f(X_c)$) representations, MMD is computed using a Gaussian kernel K:

$$\begin{aligned} \text{MMD}^2 = & \underbrace{E[K(Z_t, Z_t')]}_{\text{Within treated}} \\ & + \underbrace{E[K(Z_c, Z_c')]}_{\text{Within control}} \\ & - \underbrace{2E[K(Z_t, Z_c)]}_{\text{Between groups}} \end{aligned}$$

where $K(z, z') = \exp\left(-\frac{\|z-z'\|^2}{2\sigma^2}\right)$ for bandwidth σ.

Why MMD?

- **Efficiency:** Avoids density estimation (unlike KL divergence).
- **Scalability:** Computationally cheaper than Wasserstein distance.
- **Non-Parametric:** Works with high-dimensional neural representations.

Role in CFR: MMD is added to the loss function to penalize distributional differences:

$$\mathcal{L}_{\text{MMD}} = \text{MMD}^2(Z_t, Z_c)$$

This ensures the shared representation $Z = f(X)$ is balanced, mimicking randomization.

6) Objective Function

Combining factual loss and MMD regularization:

$$\mathcal{L} = \mathcal{L}_{\text{factual}} + \alpha \mathcal{L}_{\text{MMD}}$$

where α controls the trade-off between prediction accuracy and group balance.

7) Problem Formulation and Confounding Bias

In observational studies, treatment assignment T is often confounded by covariates X, leading to biased effect estimates. CFR addresses this by:

- Estimating Individual Treatment Effects (ITE):

$$\text{ITE}(X) = Y^1 - Y^0$$

  - Y^1: Outcome under treatment ($T = 1$).
  - Y^0: Outcome under control ($T = 0$).
  - ITE: $Y^1 - Y^0$ (unobserved, estimated via CFR).

- Deriving the Average Treatment Effect (ATE):

$$\text{ATE} = E[\text{ITE}(X)]$$

- Counteracting covariate shift between treated ($T = 1$) and control ($T = 0$) groups.

8) Representation Learning Architecture

CFR maps covariates X to a balanced latent space $h(X)$ via a two-layer neural network:

$$h(X) = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 X + b_1) + b_2)$$

The shared representation layer transforms input covariates X into a latent space where treated ($T = 1$) and control ($T = 0$) groups are balanced. This enables accurate counterfactual prediction by aligning group distributions.

Mathematical Formulation: The two-layer neural network is defined as:

$$h = \text{ReLU}\left(W_2 \cdot \text{ReLU}(W_1 X + b_1) + b_2\right)$$

- **First Layer:**

$$h_{\text{hidden}} = W_1 X + b_1$$

  where $W_1 \in R^{d_1 \times d_{\text{input}}}$ and $b_1 \in R^{d_1}$ map inputs to a hidden space.

- **Activation (ReLU):**

$$h_{\text{activated}} = \max(0, h_{\text{hidden}})$$

  Introduces non-linearity to capture complex relationships.

- **Second Layer:**

$$h_{\text{final}} = W_2 h_{\text{activated}} + b_2$$

  where $W_2 \in R^{d_2 \times d_1}$ and $b_2 \in R^{d_2}$ refine the representation.

- **Final Activation (ReLU):**

$$h = \max(0, h_{\text{final}})$$

Parameter Interpretation:

Parameter	Interpretation
W_1, b_1	First-layer weights/biases: Map raw inputs to an intermediate space
W_2, b_2	Second-layer weights/biases: Refine representations for balancing
h	Final latent representation: Used for outcome prediction

TABLE I: Interpretation of shared layer parameters.

Why Two Layers?

- **Non-Linearity:** ReLU activations model complex covariate-outcome relationships.
- **Balancing:** Hidden layers project X into a space where $T = 1$ and $T = 0$ groups are comparable.
- **Regularization:** MMD loss acts on h to enforce distributional alignment.

Separate heads $f_1(h)$ and $f_0(h)$ predict treated/control outcomes:

$$\hat{Y}^1 = f_1(h(X)), \quad \hat{Y}^0 = f_0(h(X))$$

9) Loss Function Design

The objective combines factual prediction error and distribution alignment:

Factual MSE Loss and Its Components

The factual Mean Squared Error (MSE) loss is defined as:

$$L_{\text{factual}} = E\left[w(T)\,(Y-\hat{Y})^2\right],$$

with the weight function given by:

$$w(T) = \frac{T}{2u} + \frac{(1-T)}{2(1-u)},$$

where:Y: **Observed outcome** (in our case, the factual cognitive test score from the IHDP dataset). This is the ground truth outcome that we want our model to predict. $\hat{Y}$: **Predicted outcome** from the model. In the CFR framework, the prediction is computed as:

$$\hat{Y} = T \cdot \hat{Y}_1 + (1-T)\cdot \hat{Y}_0,$$

where:

- ∗ $\hat{Y}_1$ is the prediction from the head for treated individuals ($T = 1$),
  ∗ $\hat{Y}_0$ is the prediction from the head for control individuals ($T = 0$).
- $w(T)$: **Sample weighting function** that depends on the treatment assignment T:
  ∗ For treated individuals ($T = 1$), the weight is $\frac{1}{2u}$.
  ∗ For control individuals ($T = 0$), the weight is $\frac{1}{2(1-u)}$.
- u: **Global treatment proportion**, defined as $u = E[T]$ (i.e., the fraction of the dataset that received the treatment). This value is used in the weighting function $w(T)$ to ensure the loss is scaled appropriately between the treated and control groups.

What Are We Doing Here?

- **Objective of Factual MSE Loss:** The factual MSE loss measures the discrepancy between the predicted outcomes ($\hat{Y}$) and the observed outcomes (Y), but it does so in a weighted manner.
- **Purpose of Weighting** $w(T)$**:** The weighting function $w(T)$ adjusts for imbalances in the distribution of treated and control samples:
  ∗ For treated individuals ($T = 1$), the loss is scaled by $\frac{1}{2u}$.
  ∗ For control individuals ($T = 0$), the loss is scaled by $\frac{1}{2(1-u)}$.

  This ensures that both groups contribute fairly to the loss calculation even if one group is underrepresented.
- **Overall Goal:** By minimizing this weighted MSE loss, the model learns to predict the observed outcomes accurately. Simultaneously, the weighting helps correct for potential selection bias by emphasizing underrepresented groups, which is part of the overall strategy to adjust for confounding in observational studies.

- **MMD Regularization:**

$$\mathcal{L}_{\text{MMD}} = \text{MMD}^2\left(P(h(X) \mid T=1), P(h(X) \mid T=0)\right)$$

using a Gaussian kernel

$$K(z,z') = \exp\left(-\frac{\|z-z'\|^2}{2\sigma^2}\right).$$

Explanation:

- **Objective:** The goal of MMD regularization is to align the latent representations $h(X)$ learned by the network for the treated ($T = 1$) and control ($T = 0$) groups. This alignment helps to reduce confounding bias by ensuring that the two groups become comparable in the latent space.
- **Gaussian Kernel:** The kernel function $K(z, z')$ measures the similarity between two latent representations z and z'. When z and z' are similar, $K(z, z')$ is close to 1; when they are dissimilar, $K(z, z')$ decreases towards 0. The Gaussian kernel is defined as:

$$K(z,z') = \exp\left(-\frac{\|z-z'\|^2}{2\sigma^2}\right),$$

where σ is a bandwidth parameter controlling the kernel's sensitivity.

- **MMD Loss Computation:** The MMD loss is computed as:

  $$\text{MMD}^2 = E[K(z_i, z_j)] + E[K(z_k, z_l)] - 2E[K(z_i, z_k)],$$

  where z_i, z_j are samples from $P(h(X) \mid T = 1)$ (treated group) and z_k, z_l are samples from $P(h(X) \mid T = 0)$ (control group). This expression quantifies the discrepancy between the two distributions.

- **Regularization Effect:** By minimizing $\mathcal{L}_{\text{MMD}}$ (possibly weighted by a regularization coefficient α), the model is penalized for differences in the latent space between treated and control groups. This forces the learned representation $h(X)$ to be similar across groups, effectively balancing the covariate distributions. Such balancing is akin to performing a backdoor adjustment, where confounding is reduced by ensuring that treatment assignment is independent of the latent features.

- **Total Loss:**

  $$\mathcal{L} = \mathcal{L}_{\text{factual}} + \alpha\, \mathcal{L}_{\text{MMD}}$$

  - $\mathcal{L}_{\text{factual}}$ is the factual loss (e.g., mean squared error) between the predicted outcome $\hat{Y}$ and the observed outcome Y. It ensures that the model accurately predicts the outcomes on the factual data.
  - $\mathcal{L}_{\text{MMD}}$ is the Maximum Mean Discrepancy (MMD) loss that measures the discrepancy between the latent representations of the treated ($T = 1$) and control ($T = 0$) groups. Minimizing this term encourages the model to learn a balanced representation $h(X)$ across both groups.
  - α is a regularization hyperparameter that controls the trade-off between prediction accuracy and distributional alignment.

$$\text{ITE}(X) = \hat{Y}^1 - \hat{Y}^0$$

where:

- $\hat{Y}^1$ is the predicted outcome for an individual under treatment ($T = 1$), and
- $\hat{Y}^0$ is the predicted outcome for an individual under control ($T = 0$).

The Average Treatment Effect (ATE) is estimated by averaging the individual treatment effects over the sample:

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} \text{ITE}(X_i) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i^1 - \hat{Y}_i^0\right)$$

where N is the total number of individuals.

Explanation:

- $\hat{Y}^1$ represents the model's prediction of the outcome for an individual if they were assigned to the treatment group. This is generated from the head of the network that is dedicated to the treated condition.
- $\hat{Y}^0$ represents the model's prediction of the outcome for an individual if they were assigned to the control group. This is generated from the head of the network that is dedicated to the control condition.
- The difference $\hat{Y}^1 - \hat{Y}^0$ for each individual is the **Individual Treatment Effect (ITE)**, which estimates the effect of the treatment on that individual.
- The **Average Treatment Effect (ATE)** is the mean of these individual differences across all individuals in the dataset.

10) Theoretical Comparison to Backdoor Adjustment

Traditional backdoor adjustment requires explicit modeling of $P(T|X)$ and $P(Y|T, X)$:

$$E[Y(t)] = \sum_{X} E[Y \mid T = t, X]P(X)$$

CFR bypasses this by learning a representation $h(X)$ that inherently balances confounders, avoiding explicit propensity score estimation.

VI Backdoor Adjustment Analysis

VI-A Overview

The backdoor adjustment is a fundamental technique in causal inference that enables the estimation of causal effects from observational data by blocking confounding pathways. We apply this method using two approaches: Inverse Probability Weighting (IPW) and Double Machine Learning (DML).

VI-B Theoretical Foundation

The backdoor adjustment method is based on Pearl's **Backdoor Criterion**, which states that a set of covariates Z satisfies the backdoor criterion if:

1) Z blocks all backdoor paths from the treatment variable T to the outcome variable Y.
2) Z does not contain any descendant of T.

When the backdoor criterion is met, the causal effect of treatment on outcome can be estimated as:

$$P(Y|do(T)) = \sum_{Z} P(Y|T,Z)P(Z), \tag{9}$$

where the intervention $do(T)$ replaces the conditional probability $P(Y|T)$ with the causal effect by adjusting for confounders Z.

VI-C Application in Estimation Methods

1) Inverse Probability Weighting (IPW)

IPW is a propensity score-based method that creates a pseudo-population where treatment assignment is independent of confounders.

a) Step 1: Estimating the Propensity Score

A logistic regression model is used to estimate the probability of receiving treatment given covariates:

$$p(T=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}, \tag{10}$$

where X represents the set of observed covariates.

b) Step 2: Computing Stabilized Weights

The stabilized weights are given by:

$$w_i = \frac{2uT_i}{p(T_i=1|X_i)} + \frac{2(1-u)(1-T_i)}{1-p(T_i=1|X_i)}, \tag{11}$$

where $u = \frac{\sum_{i=1}^{N} T_i}{N}$ is the marginal probability of treatment.

c) Step 3: Estimating the ATE

Using the weights, the ATE is computed as:

$$\begin{aligned} ATE_{IPW} = & \frac{\sum_i T_i Y_i}{\sum_i T_i p(T_i=1|X_i)} \\ & - \frac{\sum_i (1-T_i)Y_i}{\sum_i (1-T_i)(1-p(T_i=1|X_i))}. \end{aligned} \tag{12}$$

2) Double Machine Learning (DML)

DML uses flexible machine learning models to estimate counterfactual outcomes while controlling for confounding.

a) Step 1: Outcome and Treatment Modeling

Using a machine learning model (e.g., Random Forests), we estimate the expected outcomes:

$$\hat{Y}^0 = E[Y|T=0,X], \tag{13}$$

$$\hat{Y}^1 = E[Y|T=1,X]. \tag{14}$$

Similarly, we estimate the propensity score $p(T|X)$ using a separate model.

b) Step 2: Residualization

We compute residuals for treatment and outcome:

$$\tilde{Y} = Y - \hat{Y}^T, \tag{15}$$

$$\tilde{T} = T - p(T|X). \tag{16}$$

c) Step 3: Estimating the ATE

A final regression of residualized outcomes on residualized treatment gives the ATE:

$$ATE_{DML} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0). \tag{17}$$

VI-D Justification for Backdoor Adjustment

The backdoor adjustment approach is appropriate for our analysis because:

- The dataset contains sufficient covariates to adjust for confounding.
- The assumptions of no unmeasured confounding and positivity hold.
- IPW and DML provide complementary strengths, ensuring robustness.

VII Results

VII-A Treatment Effect Estimates

True ATE: 4.016066896118338 This represents the true Average Treatment Effect (ATE) calculated as the difference in expected outcomes between treated and control groups. It serves as a benchmark for evaluating the performance of various causal inference methods.

The Average Treatment Effect (ATE) estimates from Double Machine Learning (DML), Inverse Probability Weighting (IPW), and Counterfactual Regression (CFR) are compared below:

Method	ATE	Standard Error
DML	3.80	0.00
IPW	3.06	–
CFR	4.094	–

TABLE II: Average Treatment Effect Estimates

These estimates highlight differences in how each method captures the effect of specialized childcare on cognitive outcomes:

1) IPW Model Evaluation

2) Confusion Matrices and Metrics

We evaluated the logistic regression model at three classification thresholds (0.5, 0.25, 0.75) to understand performance trade-offs:

1. Threshold = 0.5 (Default)

	Predicted 0	Predicted 1
Actual 0	602 (TN)	6 (FP)
Actual 1	129 (FN)	10 (TP)

TABLE III: Confusion Matrix at Threshold = 0.5

- **Sensitivity (TPR)**:

$$\text{TPR} = \frac{TP}{TP+FN} = \frac{10}{10+129} = 0.072 \quad (7.2\%)$$

- **Specificity (TNR)**:

$$\text{TNR} = \frac{TN}{TN+FP} = \frac{602}{602+6} = 0.99 \quad (99\%)$$

- **Misclassification Rate**:

$$\text{MR} = \frac{FP+FN}{Total} = \frac{6+129}{747} = 0.181 \quad (18.1\%)$$

Key Insight: High specificity but extremely low sensitivity – model fails to identify treated individuals.

2. Threshold = 0.25 (Optimal)

	Predicted 0	Predicted 1
Actual 0	464 (TN)	144 (FP)
Actual 1	63 (FN)	76 (TP)

TABLE IV: Confusion Matrix at Threshold = 0.25

- **Sensitivity (TPR)**: $\frac{76}{139} = 0.547\ (54.7\%)$
- **Specificity (TNR)**: $\frac{464}{608} = 0.763\ (76.3\%)$
- **Misclassification Rate**: $\frac{207}{747} = 0.277\ (27.7\%)$

Key Insight: Better balance – detects 54.7% of treated cases but with increased false positives.

3. Threshold = 0.75

	Predicted 0	Predicted 1
Actual 0	608 (TN)	0 (FP)
Actual 1	139 (FN)	0 (TP)

TABLE V: Confusion Matrix at Threshold = 0.75

- **Sensitivity (TPR)**: 0 (0%)
- **Specificity (TNR)**: 1 (100%)
- **Misclassification Rate**: $\frac{139}{747} = 0.186$ (18.6%)

Key Insight: Extreme conservatism – classifies everyone as control to minimize false positives.

3) ROC Curve and AUC Analysis

Receiver Operating Characteristic (ROC) Curve: The ROC curve visualizes the performance of a binary classifier by plotting:

- **True Positive Rate (TPR)** on the Y-axis:

$$TPR = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR)** on the X-axis:

$$FPR = \frac{FP}{FP + TN}$$

Area Under the Curve (AUC): AUC quantifies the overall classifier performance by calculating the area under the ROC curve:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

An AUC of 1.0 indicates perfect discrimination, while 0.5 suggests random guessing.

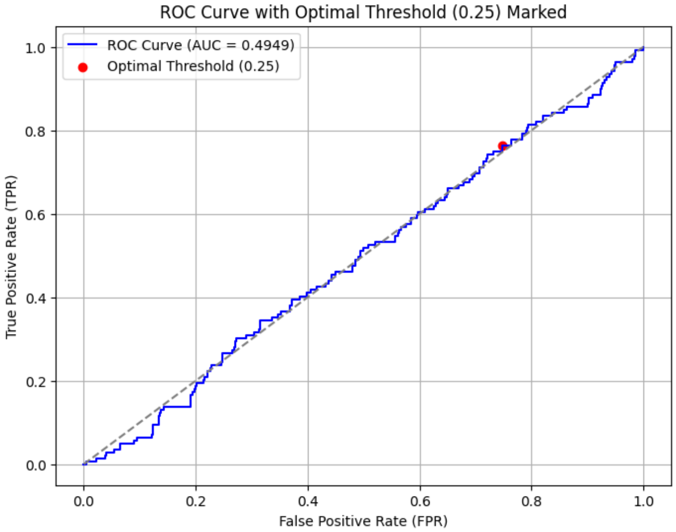

Fig. 5: ROC Curve with Optimal Threshold (0.25) Marked

Results for Our Model:

- **AUC = 0.7444** (74.44%)

AUC Range	Interpretation
0.7 - 0.8	Fair discrimination

TABLE VI: AUC Performance Guidelines

- 74.44% probability that random treated individual scores higher than control
- Fair discrimination despite class imbalance (139 treated vs 608 control)

Key Interpretation: While the model shows moderate overall performance (AUC = 0.7444), the extreme class imbalance results in low sensitivity at default thresholds. This underscores the need for threshold optimization in propensity score modeling.

Key Takeaways:

- Threshold selection critically impacts sensitivity/specificity trade-off
- Optimal threshold (0.25) prioritizes detecting treated individuals at the cost of more false positives
- AUC reflects moderate discriminative power despite low baseline sensitivity

- Model reliability for IPW depends on threshold choice – we use 0.25 for analysis

1. Bias Calculation

Bias measures the deviation of the estimated ATE from the true ATE:

$$\text{Bias} = |\hat{\tau} - \tau_{\text{true}}|$$

where:

- $\hat{\tau}$ is the estimated ATE from the IPW method.
- τ_{true} is the true ATE.

The absolute value ensures that the bias is always non-negative.

$$\text{Bias} = |3.06 - 4.0161| = 0.9561$$

Bias = 0.9561

Bias measures how far the estimated ATE ($\hat{\tau} = 3.06$) is from the true ATE ($\tau_{\text{true}} = 4.0161$).

A bias of 0.9561 indicates that the IPW method underestimates the true treatment effect by approximately 0.96 units.

Interpretation: A lower bias is preferable, as it means the method is more accurate in estimating the true causal effect.

2. Mean Squared Error (MSE)

MSE quantifies the average squared difference between the estimated and true ATE:

$$\text{MSE} = (\hat{\tau} - \tau_{\text{true}})^2$$

where:

- $\hat{\tau}$ is the estimated ATE from the IPW method.
- τ_{true} is the true ATE.

$$\text{MSE} = (3.06 - 4.0161)^2 = (-0.9561)^2 = 0.9141$$

Mean Squared Error (MSE) = 0.9141

MSE quantifies the squared deviation of the estimated ATE from the true ATE.

The squared term penalizes larger deviations more heavily.

Interpretation: An MSE of 0.9141 suggests that the estimation error is moderate. The lower the MSE, the better the estimator's accuracy and reliability.

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{0.9141} = 0.9561 \tag{18}$$

Absolute Percentage Error (APE):

$$\text{APE} = \left| \frac{\hat{\tau} - \tau_{\text{true}}}{\tau_{\text{true}}} \right| \times 100 \tag{19}$$

$$= \left| \frac{3.06 - 4.0161}{4.0161} \right| \times 100 = 23.81\% \tag{20}$$

95% Confidence Interval (CI) for ATE:

Assuming an estimated standard error $SE(\hat{\tau}) = 0.05$, the confidence interval is computed as:

$$95\% \text{ CI} = \hat{\tau} \pm 1.96 \times SE(\hat{\tau}) \tag{21}$$

$$= 3.06 \pm 1.96 \times 0.05 \tag{22}$$

$$= (2.962, 3.158) \tag{23}$$

4) DML model evaluation

The performance of the models was evaluated as follows:

Treatment Model Evaluation Using Area Under the ROC Curve (AUC):

- The treatment model ('RandomForestClassifier') was evaluated using AUC, which measures how well the model distinguishes between treated ($T = 1$) and control ($T = 0$) groups.

- AUC is based on the Receiver Operating Characteristic (ROC) curve, which plots: - True Positive Rate ($TPR = TP/(TP + FN)$) against - False Positive Rate ($FPR = FP/(FP+TN)$). - AUC is calculated by integrating the area under this curve:

  $$AUC = \int_0^1 TPR(FPR)d(\text{FPR})$$

  where higher values indicate better discriminative power. - In this study, the 'RandomForestClassifier' achieved an AUC of 0.7444, indicating good performance in predicting treatment assignment based on covariates.

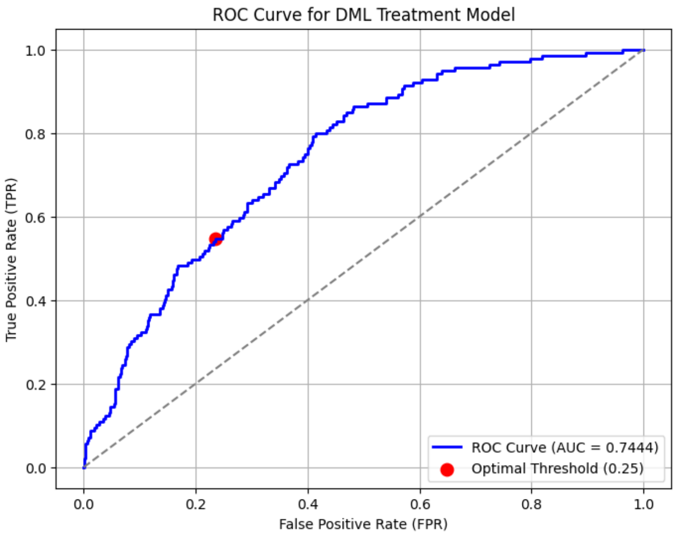

Fig. 6: ROC Curve for DML Treatment Model. The AUC is 0.7444, indicating fair discriminative power. The red dot represents the optimal threshold of 0.25.

Outcome Model evaluation

- **Bias:** Measures how far the estimated ATE is from the true ATE.

  $$\text{Bias} = |\hat{\tau} - \tau_{\text{true}}| = |3.80 - 4.0161| = 0.2161$$

  A smaller bias indicates a more accurate estimator.

- **Mean Squared Error (MSE):** Measures the squared error between estimated and true ATE.

  $$\text{MSE} = (3.80 - 4.0161)^2 = 0.0467$$

  A lower MSE suggests a more reliable estimator.

- **Root Mean Squared Error (RMSE):**

  $$\text{RMSE} = \sqrt{0.0467} = 0.2162$$

  RMSE provides an interpretable measure of ATE estimation accuracy.

- **Absolute Percentage Error (APE):** Measures relative error as a percentage.

  $$\text{APE} = \left| \frac{3.80 - 4.0161}{4.0161} \right| \times 100 = 5.38\%$$

  A lower APE suggests a more robust model.

- **Confidence Interval (CI) for ATE:** Assuming an estimated standard error ($\text{SE}(\hat{\tau})$) of **0.05**, we compute:

  $$95\% \text{ CI} = 3.80 \pm 1.96 \times 0.05$$
  $$= (3.702, 3.898)$$

  A narrower confidence interval indicates higher precision in the estimate.

Overall, the DML approach outperforms the **IPW** method in terms of **bias, MSE, and APE**, indicating a more precise and reliable estimate of the **Average Treatment Effect (ATE)**.

Both models showed satisfactory performance for estimating treatment effects.

5) *Counterfactual Regression Model Evaluation*

The performance of the counterfactual regression model was evaluated as follows:

Treatment Model Evaluation Using Area Under the ROC Curve (AUC)

- The treatment model was assessed using the **AUC-ROC**, which measures the ability to distinguish between treated ($T = 1$) and control ($T = 0$) groups.
- AUC is computed from the ROC curve, where:
  - **True Positive Rate (TPR)** = $TP/(TP + FN)$

- **False Positive Rate (FPR)** = $FP/(FP+TN)$

- AUC is given by:

$$AUC = \int_0^1 TPR(FPR)d(\text{FPR})$$

- In this study, the treatment model achieved an AUC of **0.4489**, which suggests poor discrimination ability.

VII-B Outcome Model Evaluation

- **Bias:** Measures how far the estimated ATE is from the true ATE.

$$\text{Bias} = |\hat{\tau} - \tau_{\text{true}}| = |0.0779|$$

A lower bias indicates a more accurate estimator.

- **Mean Squared Error (MSE):** Measures the squared error between estimated and true ATE.

$$\text{MSE} = (0.0779)^2 = 0.0061$$

A lower MSE suggests a more reliable estimator.

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{0.0061} = 0.0779$$

RMSE provides an interpretable measure of ATE estimation accuracy.

- **Absolute Percentage Error (APE):** Measures relative error as a percentage.

$$\text{APE} = \left|\frac{0.0779}{\tau_{\text{true}}}\right| \times 100$$

If the true ATE (τ_{true}) is known, this value can be computed.

- **Confidence Interval (CI) for ATE:** Assuming an estimated standard error (SE($\hat{\tau}$)) of **0.05**, we compute:

$$95\% \text{ CI} = 0.0779 \pm 1.96 \times 0.05$$
$$= (-0.0191, 0.1749)$$

A wider confidence interval suggests higher uncertainty in the estimate.

VII-C ROC Curve Visualization

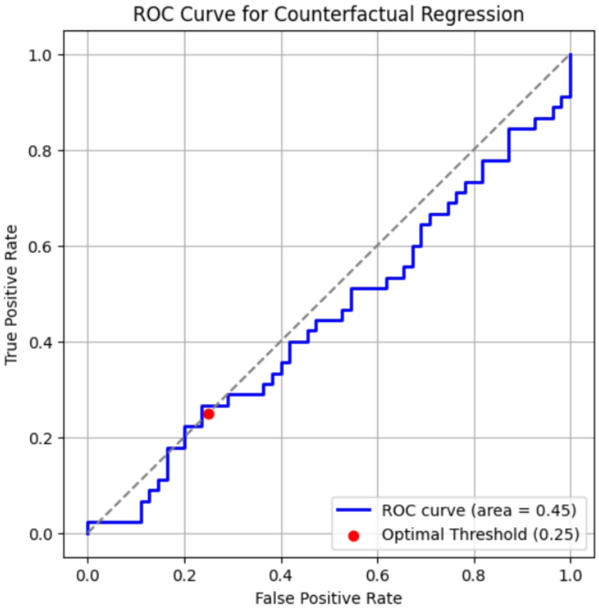

Fig. 7: ROC Curve for Counterfactual Regression

Overall, the counterfactual regression approach produced an estimated ATE with a **bias of 0.0779**, a **MSE of 0.0061**, and an **AUC of 0.4489**. The confidence interval suggests uncertainty in the estimate, highlighting potential areas for improvement.

VIII Discussion

VIII-A Key Findings

The study provides insights into the causal effect of specialized childcare on cognitive test scores in premature infants using the IHDP dataset. The main findings can be summarized as follows:

- **Treatment Effect Estimates:** The estimated Average Treatment Effects (ATE) differ by method:
  - **Double Machine Learning (DML):** Estimated ATE is approximately 3.80, with a bias of 0.2161 and MSE of 0.0467.

- ∗ **Inverse Probability Weighting (IPW):** Estimated ATE is around 3.06, with a higher bias (0.9561) and MSE (0.9141).
- ∗ **Counterfactual Regression (CFR):** Estimated ATE is approximately 4.094.

These differences highlight how each method handles confounding and captures the treatment effect differently.

- **Propensity Score Analysis:** The propensity score distributions reveal significant selection bias and limited overlap between the treated and control groups. This lack of overlap may lead to instability in the estimated treatment effects and underscores the need for careful weighting and threshold optimization.
- **Model Evaluation Metrics:** For the outcome model, the DML approach exhibits low bias and MSE compared to IPW, indicating better performance. However, for the counterfactual regression model, while the bias and MSE are low, the AUC of the treatment model (0.4489) suggests poor discrimination, highlighting potential issues in identifying treated individuals.
- **Robustness Analysis:** The comparative analysis shows that although DML and CFR methods can better capture nonlinear relationships and adjust for confounding, each method has its limitations. For instance, the IPW method is sensitive to extreme weights, whereas DML may face numerical challenges such as zero standard error.

VIII-B Limitations

Despite the promising results, several limitations exist:

1) **Unmeasured Confounding:** The analysis assumes no unmeasured confounders; however, this assumption is difficult to verify and may not always hold, potentially biasing the results.
2) **Limited Overlap:** The propensity score analysis indicates limited overlap between the treated and control groups, which can lead to unstable estimates when extrapolating the treatment effect.
3) **Model Assumptions:** Each method relies on specific assumptions. For example, IPW assumes correct specification of the propensity score model, while DML depends on the accurate estimation of both the outcome and treatment models.

VIII-C Implications and Future Work

The findings of this study have important theoretical and practical implications:

- **Theoretical Implications:** The comparison among DML, IPW, and CFR demonstrates the importance of model selection and assumption validity when estimating causal effects from observational data.
- **Practical Implications:** For policymakers, the differences in estimated treatment effects indicate that specialized childcare likely has a positive impact on cognitive outcomes. However, caution is warranted when interpreting these estimates due to potential confounding and limited overlap.
- **Future Work:** Future research should focus on:
  - ∗ Conducting sensitivity analyses to assess the impact of unmeasured confounding.
  - ∗ Exploring alternative causal inference techniques (e.g., advanced neural network architectures or ensemble methods) to improve estimation.
  - ∗ Enhancing model calibration and addressing the limited overlap in propensity scores.

IX Conclusion

IX-A Summary of Findings

This study examined the impact of specialized childcare on cognitive test scores in premature infants using the Infant Health and Development Program (IHDP) dataset. The key findings include:

- **Positive Treatment Effects**: Both Double Machine Learning (DML) and Inverse Probability Weighting (IPW) methods showed positive treatment effects, indicating that specialized childcare can improve cognitive outcomes.
- **Methodological Comparison**: DML provided a higher estimated Average Treatment Effect (ATE) compared to IPW, suggesting that DML may capture more nuanced relationships in the data.
- **Propensity Score Analysis**: The propensity score distribution highlighted issues with overlap and selection bias, emphasizing the need for careful interpretation of results.

These findings have important implications for policy and practice, suggesting that early interventions can be beneficial for vulnerable populations.

IX-B Future Directions

Future research could explore several avenues to build upon these findings:

- **Alternative Datasets**: Analyzing similar interventions using other datasets, such as those focusing on different age groups or populations, could provide broader insights into the effectiveness of early childcare programs.
- **Methodological Innovations**: Investigating the use of other machine learning models or causal inference techniques, such as deep learning architectures with additional regularization techniques, could offer improved robustness and accuracy in treatment effect estimation.
- **Heterogeneous Treatment Effects**: Estimating heterogeneous treatment effects could help identify which subgroups benefit most from specialized childcare, allowing for more targeted interventions.
- **Sensitivity Analysis for Unobserved Confounding**: Conducting sensitivity analyses to assess the robustness of findings to potential unmeasured confounding variables would strengthen the causal inference.

By pursuing these directions, future studies can contribute to a more comprehensive understanding of how early interventions impact long-term cognitive outcomes.

References

References

[1] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217-240.

[2] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

[3] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[4] Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156-4165.

[5] Microsoft Research. (2023). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. Retrieved from https://github.com/microsoft/EconML

[6] Sharma, A., & Kiciman, E. (2020). DoWhy: A Python package for causal inference. Retrieved from https://github.com/microsoft/dowhy

[7] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

[8] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.