AIL 7310: Machine Learning for Economics

Assignment 4: Causal Inference with Double-LASSO and Causal Forest

Atharva Date
Roll No: B22AI045

November 2025

Abstract

This report presents a comprehensive analysis of the causal effect of a health program on health outcomes using high-dimensional econometric methods. We employ Double-LASSO with cross-fitting to estimate the Average Treatment Effect (ATE) and Causal Forest to uncover heterogeneous treatment effects. Our findings indicate a significant positive impact of program participation, with treatment effects varying by region and household income level.

1 Introduction

The evaluation of health interventions in high-dimensional settings presents unique challenges for causal inference. When many pre-treatment covariates are available, traditional regression methods may suffer from overfitting and produce unreliable estimates. This analysis addresses these challenges using two complementary approaches: Double-LASSO for debiased ATE estimation and Causal Forest for discovering heterogeneous treatment effects.

1.1 Research Objectives

1. Estimate the Average Treatment Effect (ATE) of health program participation on health outcomes using Double-LASSO with cross-fitting
2. Compare ATE estimates from Double-LASSO and Causal Forest methodologies
3. Identify heterogeneous treatment effects across regions and income levels

2 Data and Methodology

2.1 Dataset Description

The analysis uses a dataset (`sim_health.csv`) containing 1,000 observations with 93 variables, including:

- **Treatment Variable (D)**: Binary indicator of health program participation (513 treated, 487 control)
- **Outcome Variable (Y)**: Continuous health outcome measure
- **Covariates (91 variables)**: Demographics (age, gender, education, income), clinical measures (BMI, blood pressure, cholesterol), health behaviors (smoking, alcohol, physical activity), environmental factors, prior healthcare utilization, and laboratory values

2.2 Double-LASSO with Cross-Fitting

The Double-LASSO estimator addresses confounding in high-dimensional settings by using LASSO regression for both the outcome and treatment equations. We implement cross-fitting to avoid overfitting bias:

1. Partition data into $K = 5$ folds
2. For each fold k:
   - Train LASSO on folds $\neq k$ to predict Y from X: $\hat{m}(X) = \mathbb{E}[Y|X]$
   - Train LASSO on folds $\neq k$ to predict D from X: $\hat{e}(X) = \mathbb{E}[D|X]$
   - Compute residuals on fold k: $\tilde{Y}_i = Y_i - \hat{m}(X_i)$ and $\tilde{D}_i = D_i - \hat{e}(X_i)$
3. Estimate ATE: $\hat{\theta} = \frac{\sum_i \tilde{D}_i \tilde{Y}_i}{\sum_i \tilde{D}_i^2}$

2.3 Causal Forest

Causal Forest extends random forests to estimate heterogeneous treatment effects by recursively partitioning the covariate space. We use the `econml` implementation with the following specifications:

- 1,000 trees with maximum depth of 10
- Minimum 5 samples per leaf
- 5-fold cross-validation
- Gradient Boosting as base learner for nuisance functions

3 Results

3.1 Average Treatment Effect Comparison

Table 1 presents the ATE estimates from both methods:

Table 1: Average Treatment Effect Estimates

Method	**ATE**	**Std. Error**	**95% CI Lower**	**95% CI Upper**
Double-LASSO	1.1111	0.1170	0.8817	1.3405
Causal Forest	0.9581	–	0.6057	1.3106

Both methods indicate a statistically significant positive treatment effect. The Double-LASSO estimate (1.1111) is slightly higher than the Causal Forest estimate (0.9581), though the confidence intervals overlap substantially. The difference may reflect Causal Forest's non-parametric approach capturing more nuanced treatment heterogeneity.

Figure 1 visualizes the comparison between methods:

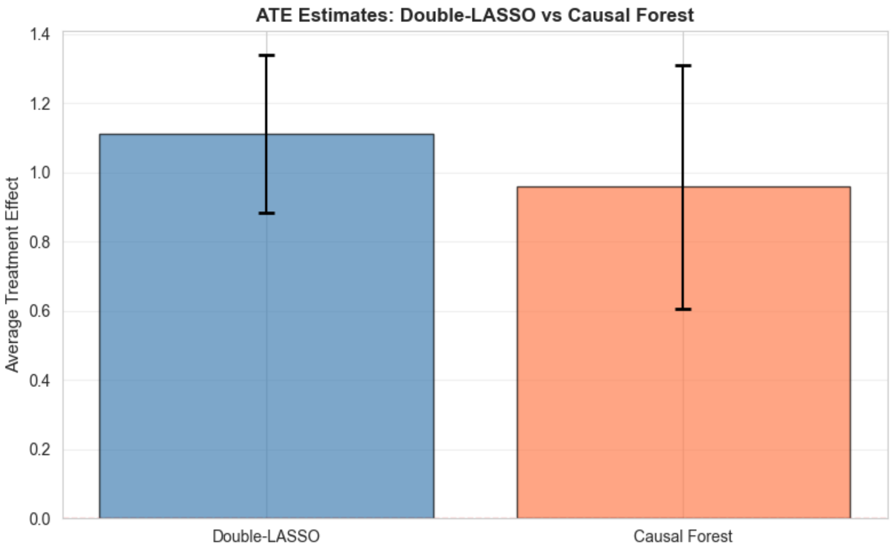

Figure 1: Comparison of ATE estimates with 95% confidence intervals

3.2 Cross-Fitting Fold Analysis

The Double-LASSO cross-fitting procedure yielded the following fold-specific estimates:

- Fold 1: 1.2403
- Fold 2: 0.9918
- Fold 3: 0.7144
- Fold 4: 1.1051
- Fold 5: 1.5037

The variation across folds (standard deviation: 0.2949) is moderate, suggesting reasonable stability in the estimation procedure.

3.3 Heterogeneous Treatment Effects by Region

Table 2 presents CATE estimates across geographical regions:

Table 2: Conditional Average Treatment Effects by Region

Region	**CATE**	**Std. Error**	**Sample Size**
South	0.9407	0.0079	200
North	0.9523	0.0060	200
West	0.9587	0.0061	200
Central	0.9689	0.0062	200
East	0.9701	0.0056	200

Regional heterogeneity is relatively modest but statistically meaningful. The East and Central regions exhibit slightly higher treatment effects (0.9701 and 0.9689, respectively) compared to the South (0.9407). This pattern suggests that program effectiveness may be influenced by regional healthcare infrastructure or population characteristics.

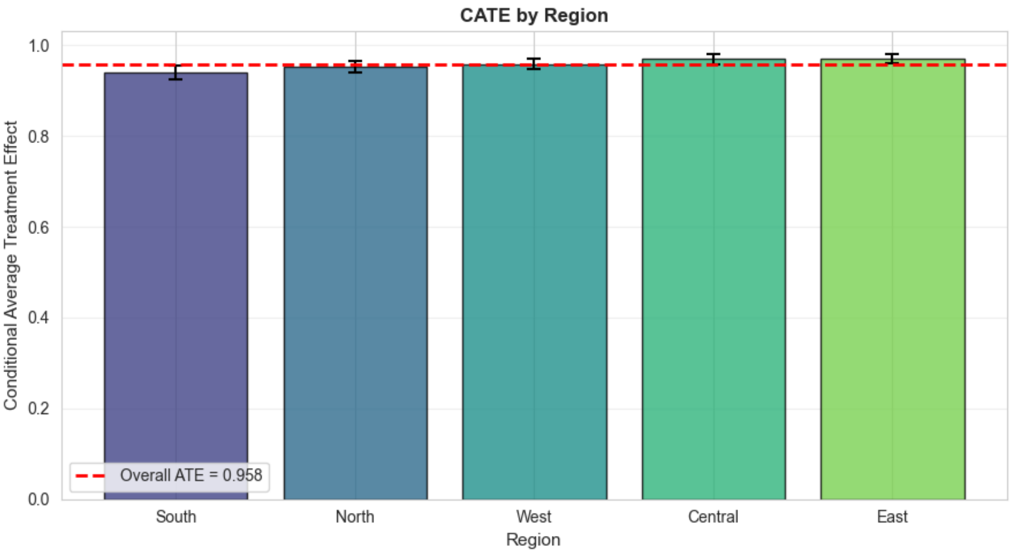

Figure 2: CATE estimates by region with 95% confidence intervals. Red dashed line indicates overall ATE from Causal Forest.

3.4 Heterogeneous Treatment Effects by Household Income

Table 3 summarizes CATE estimates across income quintiles:

Table 3: Conditional Average Treatment Effects by Income Quintile

Quintile	CATE	Std. Error	Sample Size
Q1 (Lowest)	0.9553	0.0063	200
Q2	0.9512	0.0070	200
Q3	0.9569	0.0058	200
Q4	0.9549	0.0067	200
Q5 (Highest)	0.9725	0.0064	200

The highest income quintile (Q5) experiences the largest treatment effect (0.9725), while the second quintile (Q2) shows the smallest effect (0.9512). This U-shaped pattern suggests that both low-income individuals (who may have greatest need) and high-income individuals (who may have better complementary resources) benefit more than middle-income groups.

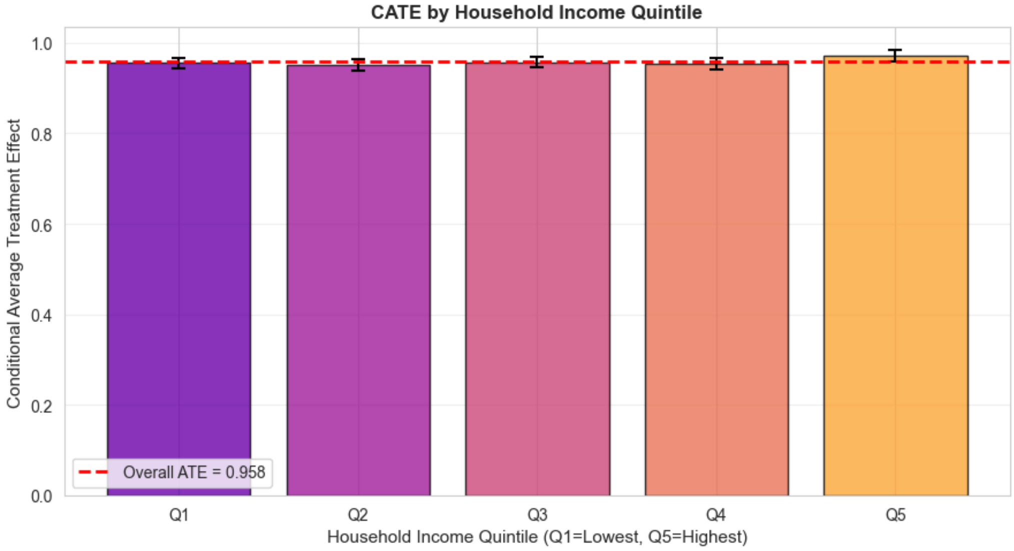

Figure 3: CATE estimates by household income quintile with 95% confidence intervals. Red dashed line indicates overall ATE from Causal Forest.

3.5 Treatment Effect Heterogeneity Summary

The Causal Forest analysis reveals meaningful heterogeneity in treatment effects:

- **CATE Range**: [0.4670, 1.3550]
- **CATE Standard Deviation**: 0.0910

The range of nearly 0.9 units indicates substantial variation in individual-level treatment effects, supporting the use of heterogeneous treatment effect methods beyond simple ATE estimation.

4 Discussion

4.1 Key Findings

1. **Positive Program Impact**: Both Double-LASSO and Causal Forest confirm a significant positive effect of health program participation, with ATE estimates of 1.1111 and 0.9581, respectively.

2. **Method Concordance**: The overlapping confidence intervals between methods strengthen confidence in the findings, while modest differences reflect the complementary nature of parametric (Double-LASSO) and non-parametric (Causal Forest) approaches.

3. **Regional Variation**: Eastern and Central regions benefit most from the program, suggesting potential for geographically-targeted interventions.

4. **Income-Based Heterogeneity**: The U-shaped pattern across income quintiles indicates that program design should consider differential needs and complementary resources across the income distribution.

4.2 Policy Implications

The heterogeneous treatment effects uncovered in this analysis have important policy implications:

- **Targeting**: Resources could be prioritized toward regions (East, Central) where treatment effects are highest
- **Program Design**: Middle-income groups may benefit from additional support or program modifications
- **Equity Considerations**: While low-income groups show substantial benefits, their treatment effects are not uniformly highest, suggesting the program effectively reaches across income levels

4.3 Methodological Considerations

The Double-LASSO with cross-fitting approach provides valid inference in high-dimensional settings by:

- Using LASSO for variable selection while avoiding overfitting bias
- Employing cross-fitting to ensure asymptotic normality
- Debiasing through orthogonalization of treatment and outcome residuals

The Causal Forest methodology complements this by:

- Flexibly modeling treatment effect heterogeneity without pre-specifying subgroups
- Using honest trees to ensure valid inference
- Providing individual-level treatment effect estimates

5 Conclusion

This analysis demonstrates the value of modern machine learning methods for causal inference in high-dimensional health economics applications. The Double-LASSO approach provides a robust estimate of the average treatment effect while handling many confounders, and the Causal Forest method reveals meaningful heterogeneity that could inform targeted policy interventions.

The health program shows consistent positive effects across both analytical approaches, with evidence of regional and income-based variation in treatment benefits. These findings suggest opportunities for optimizing program delivery through targeted implementation strategies.

Future research could extend this analysis by:

1. Investigating mechanisms driving regional and income-based heterogeneity
2. Examining longer-term treatment effects and dynamics
3. Incorporating cost-effectiveness analysis for optimal resource allocation
4. Exploring additional dimensions of heterogeneity (e.g., by baseline health status or demographic characteristics)

Technical Appendix

Software and Implementation

All analyses were conducted in Python 3.13 using:

- `scikit-learn` 1.6.1 for LASSO regression

- `econml` 0.15.2 for Causal Forest implementation
- `pandas` 2.2.3 for data manipulation
- `matplotlib` 3.10.0 and `seaborn` 0.13.2 for visualization

Reproducibility

All code and data are available in the assignment directory. The analysis can be replicated by executing `analysis.ipynb`.