Machine Learning in Mediation Analysis: An Extension of Inverse Probability Weighting

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting

Eliza, Krystal, Lily, Shu Ting

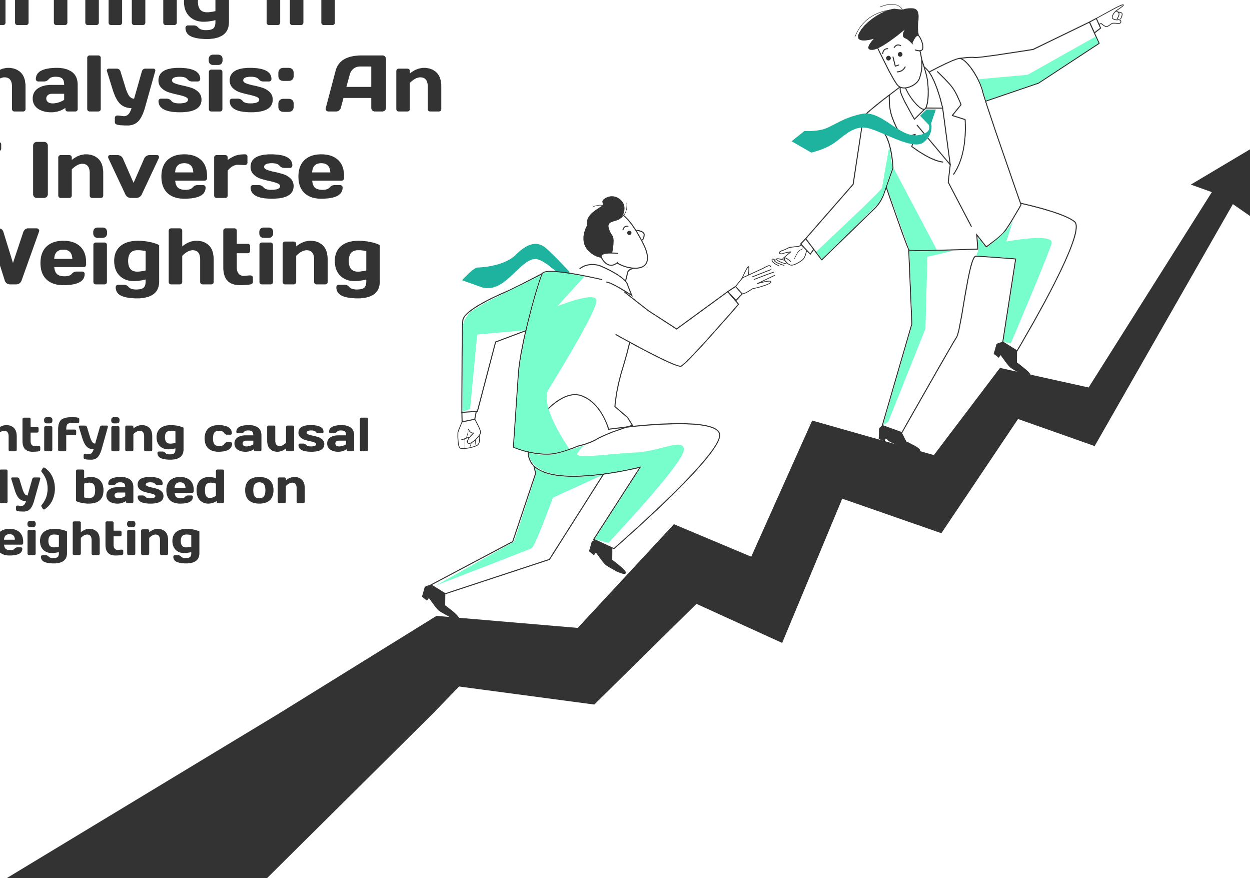


Table of contents

01 Motivation and Methodology

02 Simulation

03 Results

04 Discussion

Motivation

- We want to estimate the total effect contributed by mediators.
  - The evaluation of direct and indirect effects is referred to as **mediation analysis.**

- Challenge: randomness of treatment assignment does not guarantee that the mediator is random.

Inverse Probability Weighting

- Units are weighted by the inverse of their conditional propensity to be observed in a particular treatment state given the mediator (M) and the observed covariates (X)

- Corrects for confounding in observational studies – where treatment isn't randomised – by weighting on the predicted probability of receiving a treatment.

Assumptions

Random treatment assignment and conditional independence

Unconfoundedness of the treatment on the mediator and outcome must hold when conditioned on observed covariates.

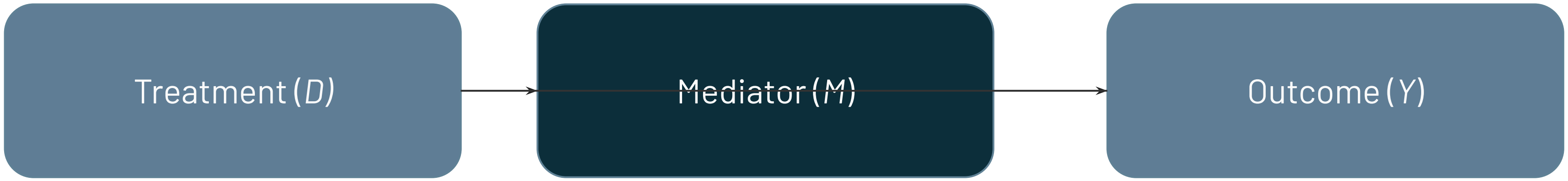
Functional Form restriction w.r.t potential mediators

The relationship between Y and M given the treatment and potential mediator M(d) is the same functional form, regardless of the treatment state of M.

Common Support

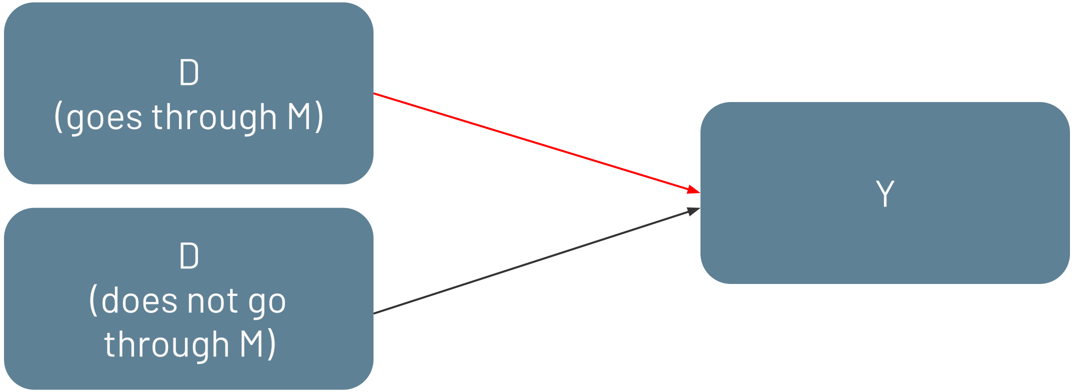
For every combination of mediator and observed covariates, there is always a mix of treated and untreated units.

Direct Effect



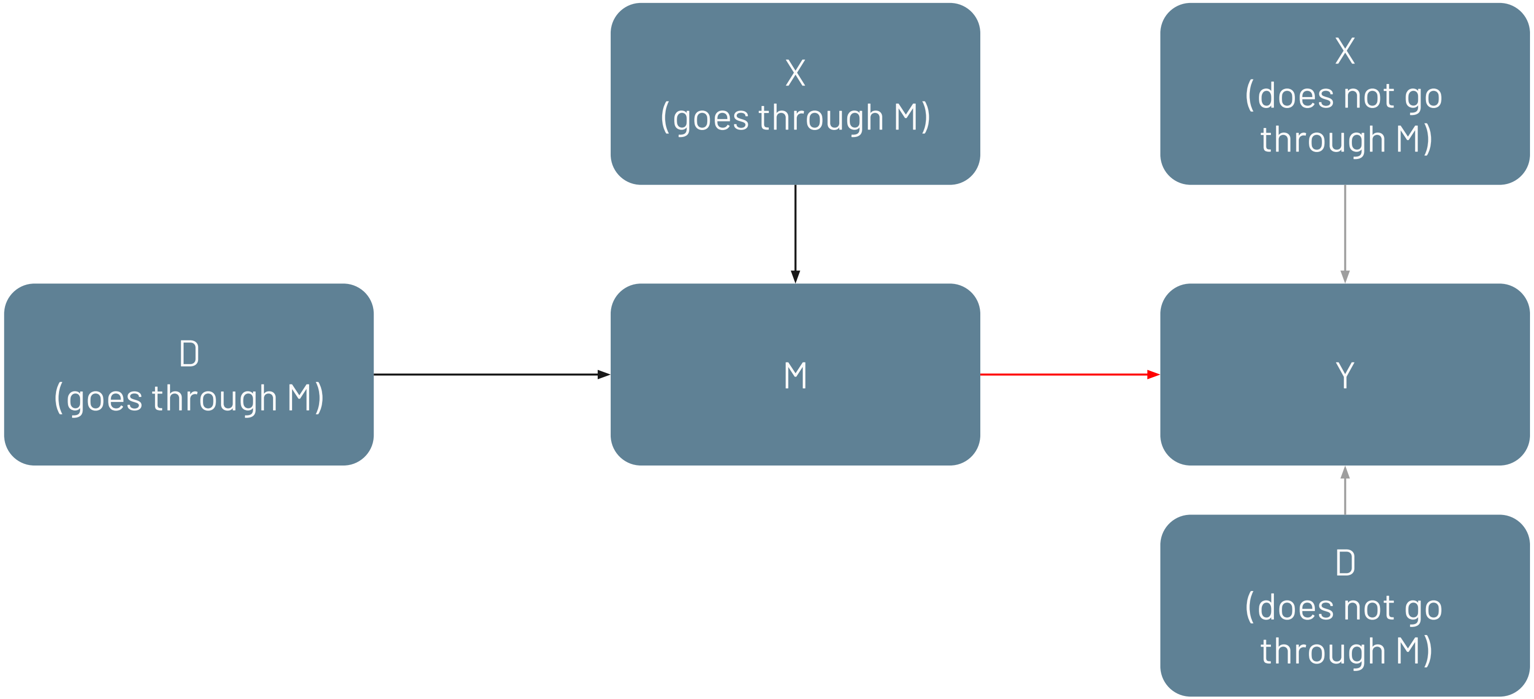
- Treatment effect while holding the mediator constant
- Reweighted by the probability of getting treatment given the mediator and covariates.

Partial Indirect Effect



- The effect from D going through M

Total Indirect Effect

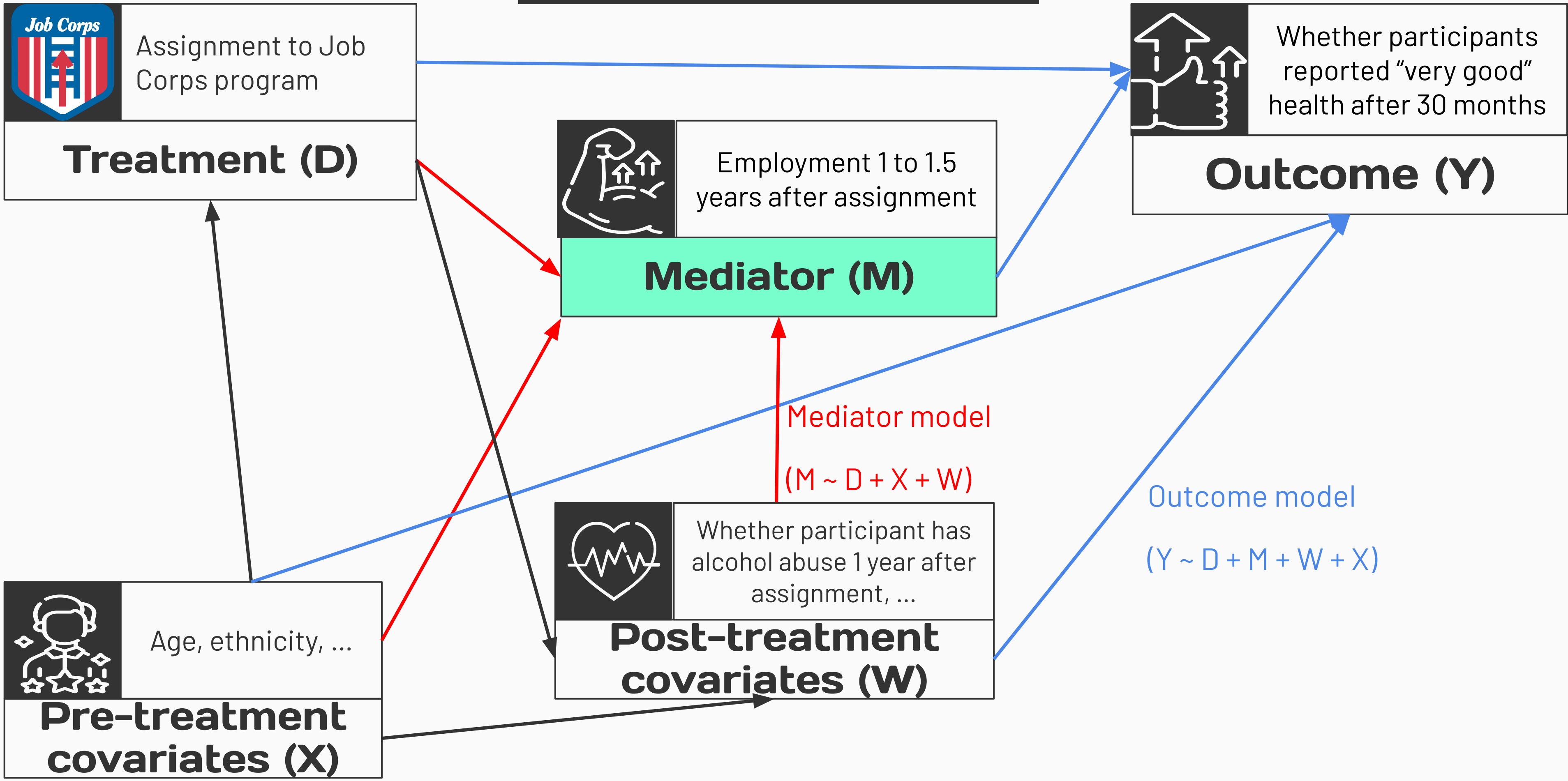


- All effects via M which either come from D or through X

Job Corps program

- Publicly funded program in the US
- **Target audience:** low-income youth
- Provide free educational and vocational training
- **Goal:** help participants secure better job opportunities
- **Ultimate aim:** improve their quality of life

Causal Mediation



Inverse Probability Weighting

- Estimate the propensity scores using the **probit** model
- Weight **trimming** to remove propensity scores close to the boundaries of 0 and 1
- **Total** effect: E[Y(1) - Y(0)]
- **Direct** effect: $E[(\frac{Y \cdot D}{Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-Pr(D=1|M,X)}) \cdot \frac{Pr(D=d|M,X)}{Pr(D=d)}]$
- **Indirect** effect: $E[\frac{Y-\mu_{d,x}(E[M|D=1-d]) \cdot I\{D=d\}}{Pr(D=d)}]$

Double Machine Learning

- 1 part of the data is used for estimating the **model parameters** using post LASSO regression with default settings
- 1 part of the data is used for predicting the **efficient score functions**
- **Cross fitting** - swap the roles of the 2 data parts
- Take **average** of the predicted efficient score functions in the combined sample to get direct & indirect effects

Doubly Robust

- **Propensity score model** (IPW): estimate $\Pr(D = 1 \mid M, X)$
- **Outcome model** (regression): $Y \sim D + M$
- Doubly robust estimator **combines** outcomes weighted using IPW with predictions generated by outcome model to estimate direct & indirect effects

Causal Forest

- Ensemble of honest causal trees trained via **subsampling** & **EMSE-optimized splits**
- Partition data into subgroups with **distinct** treatment effects
- **Decompose** total effect into direct & indirect using
  - Total effect: (Y ~ D | X)
  - Direct effect: (M ~ D | X) → get M hat → (Y ~ D + M hat | X)
  - Indirect effect: (Y ~ M | X)

Results

Gender	Model	Effect	Estimate	Confidence Interval
Female	Inverse Probability Weighting	Total	0.0285	(0.00003, 0.05796)
		Direct	0.0307	(-0.00323, 0.06522)
		Indirect	0.00195	(-0.00906, 0.01326)
	Double Machine Learning	Total	0.0261	(-0.0058, 0.058)
		Direct	0.0279	(-0.0039, 0.0596)
		Indirect	0.00149	(0.0001, 0.00288)
	Doubly Robust	Total	0.0165	(-0.00100, 0.05885)
		Direct	0.0282	(-0.00088, 0.05847)
		Indirect	-0.0117	(-0.00092, 0.00150)
	Causal Forest	Total	0.0262	(0.02442, 0.02762)
		Direct	0.0264	(0.0247, 0.02784)
		Indirect	0.00465	(0.00282, 0.00603)
	Causal Mediation	Total	0.0258	(-0.00381, 0.05)
		Direct	0.0259	(-0.00409, 0.05)
		Indirect	0.00119	(-0.00097, 0)

Gender	Model	Effect	Estimate	Confidence Interval
Male	Inverse Probability Weighting	Total	0.0219	(-0.00498, 0.04717)
		Direct	0.00227	(-0.03683, 0.03891)
		Indirect	0.0173	(0.00549, 0.0296)
	Double Machine Learning	Total	0.00100	(-0.0255, 0.0343)
		Direct	0.00258	(-0.0222, 0.0391)
		Indirect	-0.00115	(-0.00169, 0.0004)
	Doubly Robust	Total	0.01084	(-0.02917, 0.03101)
		Direct	0.0005	(-0.02955, 0.03076)
		Indirect	0.0103	(-0.00075, 0.00123)
	Causal Forest	Total	0.0182	(0.01676, 0.01964)
		Direct	0.0196	(0.0186, 0.02151)
		Indirect	0.0140	(0.01268, 0.01556)
	Causal Mediation	Total	0.00299	(-0.02183, 0.03)
		Direct	0.00285	(-0.02173, 0.03)
		Indirect	0.000142	(-0.00044, 0)

Discussion

Strengths & Limitations

Addresses Critical Gap in Mediation Analysis

Flexible, avoids imposing strong functional assumptions

Corrects for mediator endogeneity

Partially addresses unobserved confounding

Partial correction of endogeneity

Susceptible to compounding misspecification

No reliable framework for indirect effect estimation

Incomplete causal conclusions

Dependence on accurate estimations & strong assumptions

Misspecification leads to unstable weights and inconsistent estimates

Violation of Common Support assumption results in inflated variance

Inverse Probability Weighting

1. Inflated standard errors

	Female					Male				
Effect	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation
ATE	0.0285 (0.0148)	0.0261 (0.0162)	0.0165 (0.0371)	0.0262 (0.0150)	0.0258 (0.0137)	0.0219 (0.0132)	0.00100 (0.0152)	0.01084 (0.0312)	0.0182 (0.0134)	0.00299 (0.0132)
Direct effect under treatment	0.0307 (0.0168)	0.0279 (0.0162)	0.0282 (0.0159)	0.0264 (1.845)	0.0259 (0.0138)	0.00227 (0.0187)	0.00258 (0.0153)	0.0005 (0.0150)	0.0196 (9.348)	0.00285 (0.0132)
Indirect effect under treatment	0.00195 (0.00531)	0.00149 (0.000710)	-0.0117 (0.7272)	0.00465 (0.0153)	0.00119 (0.0130)	0.0173 (0.00593)	-0.00115 (0.000742)	0.0103 (0.0273)	0.0140 (0.0148)	0.000142 (0.000114)

Inverse Probability Weighting

Gender	Model	Effect	MSE	Confidence Interval
Female	Inverse Probability Weighting	Total	0.00108	(3e-05, 0.05796)
		Direct	0.00125	(-0.00323, 0.06522)
		Indirect	0.00004	(-0.00906, 0.01326)
	Double Machine Learning	Total	0.00026	(-0.0058, 0.058)
		Direct	0.00026	(-0.0039, 0.0596)
		Indirect	0.00000	(1e-04, 0.00288)
	Doubly Robust	Total	0.00038	(-0.00100, 0.05885)
		Direct	0.00024	(-0.00088, 0.05847)
		Indirect	0.00014	(-0.00092, 0.00150)
	Causal Forest	Total	0.31793	(0.02442, 0.02762)
		Direct	0.31768	(0.0247, 0.02784)
		Indirect	0.33028	(0.00282, 0.00603)
	Normal Mediation	Total	0.00378	(-0.00381, 0.05)
		Direct	0.00383	(-0.00409, 0.05)
		Indirect	0.00010	(-0.00097, 0)

1. Inflated standard errors

2. Wider confidence intervals

High uncertainty

Doubly Robust

weighting + outcome modelling

only require correct specification of one model

More stable estimates for direct effects

	Female					Male				
Effect	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation
ATE	0.0285 (0.0148)	0.0261 (0.0162)	0.0165 (0.0371)	0.0262 (0.0150)	0.0258 (0.0137)	0.0219 (0.0132)	0.00100 (0.0152)	0.01084 (0.0312)	0.0182 (0.0134)	0.00299 (0.0132)
Direct effect under treatment	0.0307 (0.0168)	0.0279 (0.0162)	0.0282 (0.0159)	0.0264 (1.845)	0.0259 (0.0138)	0.00227 (0.0187)	0.00258 (0.0153)	0.0005 (0.0150)	0.0196 (9.348)	0.00285 (0.0132)
Indirect effect under treatment	0.00195 (0.00531)	0.00149 (0.000710)	-0.0117 (0.7272)	0.00465 (0.0153)	0.00119 (0.0130)	0.0173 (0.00593)	-0.00115 (0.000742)	0.0103 (0.0273)	0.0140 (0.0148)	0.000142 (0.000114)

Doubly Robust

Model	Effect	MSE	Confidence Interval
Inverse Probability Weighting	Total	0.00037	(-0.00498, 0.04717)
	Direct	0.00035	(-0.03683, 0.03891)
	Indirect	0.00033	(0.00549, 0.0296)
Doubly Robust	Total	0.00033	(-0.02917, 0.03101)
	Direct	0.00023	(-0.02955, 0.03076)
	Indirect	0.00010	(-0.00075, 0.00123)

More stable estimates for direct effects

Tighter confidence intervals + Lower MSE

Less sensitive to extreme weights/noise

Double Machine Learning

IPW

P(D=d | M,X)

DML

POST LASSO

Handle complex interactions between outcome and covariates

Cross Fitting

Reduces bias from estimation of nuisance parameters

Females:	Total MSE	Direct MSE	Indirect MSE
IPW	0.00108	0.00125	0.00004
DML	0.00026	0.00126	0.00000
CF	0.31793	0.31768	0.33028

Causal Forest

Tree Based Ensemble Approach → Subgroup Specific Mediation Analysis → Treatment Heterogeneity

	Direct Effect	Indirect Effect
Male	0.0196	0.0140
Female	0.0264	0.00465

Causal Forest

Overfitting

	Total MSE	Direct MSE	Indirect MSE
Male	0.40943	0.40789	0.41392
Female	0.31793	0.31768	0.33028

Inability to produce trimmed effects

- Sensitivity to outliers
- Trimming breaks key requirements for CF (removed portions might be critical in constructing splits)

Baseline: Causal Mediation

Strong Linear Assumptions

No Interaction Terms

Gender	Model	Effect	MSE
Female	Inverse Probability Weighting	Total	0.00108
		Direct	0.00125
		Indirect	0.00004
	Double Machine Learning	Total	0.00026
		Direct	0.00026
		Indirect	0.00000
	Doubly Robust	Total	0.00038
		Direct	0.00024
		Indirect	0.00014
	Causal Forest	Total	0.31793
		Direct	0.31768
		Indirect	0.33028
	Normal Mediation	Total	0.00378
		Direct	0.00383
		Indirect	0.00010

Extension of Huber’s IPW

Double Machine Learning

- Improved robustness and precision

Causal Forest

- Surfaces treatment heterogeneity

Doubly Robust

- Greater stability and tighter CI

Limitations

Potential Unmodelled Effect Modifiers

Potential Unobserved Covariates

	Female Total MSE	Male Total MSE
IPW	0.00108	0.00037
DML	0.00026	0.00287
DR	0.00038	0.00033
CF	0.31793	0.40943
Normal Mediation	0.00378	0.00342

Thank You

Appendix

Double Machine Learning

- The average indirect effect of the binary treatment &the unmediated direct effect are estimated based on efficient score functions, which are robust with respect to misspecifications of the outcome, mediator, and treatment models.
- Efficient score functions of the potential outcomes is slightly different from the propensity score functions of IPW. (see next page)
- **Standard errors** are based on asymptotic approximations using the estimated variance of the efficient score functions

DML VS IPW

DML:
Efficient score function

LEMMA 3.3. *Under Assumptions 3.1, 3.2 and 3.3, the potential outcome* $E[Y(d, M(d))]$ *is identified by the following efficient score function:*

$$E[Y(d, M(d))] = E[\alpha_d] \text{ with } \alpha_d = \frac{I\{D = d\} \cdot [Y - \mu(d, X)]}{p_d(X)} + \mu(d, X), \tag{3.3}$$

where $\mu(D, X) = E(Y|D, M(D), X) = E(Y|D, X)$ *is the conditional expectation of outcome* Y *given* D *and* X.

IPW:
Propensity score function

$$\begin{aligned} E[Y(d,M(d))] &= E[E[Y(d,M(d))|X=x]] = E[E[Y|D=d,X=x]] \\ &= E\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X)}|X=x\right]\right] = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X)}\right] \end{aligned}$$