A Unified Theory of Mediation Analysis: From Traditional Methods to Modern Machine Learning

Mediation Analysis Framework Comparison Project

June 28, 2025

Abstract

This document provides a comprehensive theoretical foundation for mediation analysis, spanning from traditional linear methods to modern machine learning approaches. We derive the mathematical relationships between four major frameworks: Traditional (Baron & Kenny), Frisch-Waugh-Lovell (FWL), Double Machine Learning (DML), and Causal Mediation (Natural Effects). We prove their equivalences under specific conditions and demonstrate when each approach is most appropriate. Special attention is given to the Percentage of Mediated Accuracy (PoMA) and its stability issues in edge cases, particularly symmetric non-linear relationships where traditional formulas can produce nonsensical results.

Contents

1 Introduction

Mediation analysis seeks to decompose the total effect of a treatment X on an outcome Y into direct and indirect pathways through a mediator M. The fundamental question is: "How much of the effect of X on Y operates through M?"

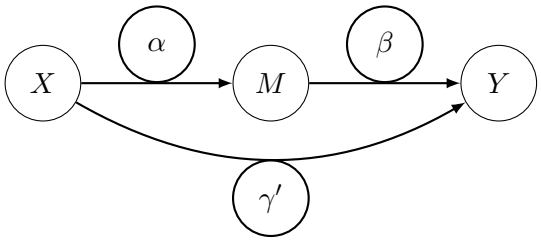

Figure 1: The classical mediation diagram

The key quantities of interest are:

- **Total Effect**: $c = \mathbb{E}[Y|X = x + 1] - \mathbb{E}[Y|X = x]$
- **Direct Effect**: $c' = \mathbb{E}[Y|X = x + 1, M = m] - \mathbb{E}[Y|X = x, M = m]$
- **Indirect Effect**: $ab = c - c'$
- **Percentage of Mediation (PoMA)**: $\text{PoMA} = \frac{ab}{c} = 1 - \frac{c'}{c}$

2 Traditional Mediation Analysis

2.1 The Baron & Kenny Approach

The traditional approach (Baron and Kenny, 1986) involves fitting two regression models:

$$Y = cX + \epsilon_1 \tag{1}$$
$$Y = c'X + \beta M + \epsilon_2 \tag{2}$$

From these regressions:

- c estimates the total effect
- c' estimates the direct effect
- β estimates the effect of M on Y controlling for X

Additionally, to estimate the path $X \to M$:

$$M = \alpha X + \epsilon_3 \tag{3}$$

Proposition 1 (Product of Coefficients)**.** *Under linearity assumptions, the indirect effect equals $\alpha\beta$, where α is from equation* (3) *and β is from equation* (2)*.*

Proof. By the linearity of expectations and assuming no unmeasured confounding:

$$\text{Total Effect} = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] = c \tag{4}$$

$$\text{Direct Effect} = \mathbb{E}[Y|X=1, M=m] - \mathbb{E}[Y|X=0, M=m] = c' \tag{5}$$

$$\text{Indirect Effect} = c - c' = \alpha\beta \tag{6}$$

□

2.2 Limitations of Traditional Approach

The traditional approach assumes:

1. Linear relationships throughout
2. No treatment-mediator interaction
3. No unmeasured confounding
4. Correctly specified models

3 Frisch-Waugh-Lovell (FWL) Approach

3.1 The FWL Theorem

The Frisch-Waugh-Lovell theorem provides an alternative way to compute partial regression coefficients through residualization.

Theorem 2 (Frisch-Waugh-Lovell)**.** *Consider the regression $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$. The coefficient β_1 can be obtained by:*

1. *Regressing Y on X_2 to get residuals $\tilde{Y}$*
2. *Regressing X_1 on X_2 to get residuals $\tilde{X}_1$*
3. *Regressing $\tilde{Y}$ on $\tilde{X}_1$*

3.2 Application to Mediation

For mediation analysis, we apply FWL to estimate the direct effect c':

Proposition 3 (FWL equals Traditional)**.** *The direct effect estimated by FWL equals the direct effect from traditional regression.*

Algorithm 1 FWL Mediation Analysis

Input: Data (X, M, Y)
Step 1: $e_Y \leftarrow Y - \mathbb{E}[Y|M]$ (residualize Y on M)
Step 2: $e_X \leftarrow X - \mathbb{E}[X|M]$ (residualize X on M)
Step 3: $c' \leftarrow \text{Cov}(e_Y, e_X)/\text{Var}(e_X)$
Return: Direct effect c'

Proof. Let $P_M = M(M'M)^{-1}M'$ be the projection matrix onto the column space of M, and let $M_\perp = I - P_M$ be the residual maker. Then:

From equation (2), the normal equations give:

$$X'Y = c'X'X + \beta X'M \tag{7}$$
$$M'Y = c'M'X + \beta M'M \tag{8}$$

Solving for c' by eliminating β:

$$c' = \frac{X'M_\perp Y}{X'M_\perp X} = \frac{e_X' e_Y}{e_X' e_X} \tag{9}$$

where $e_Y = M_\perp Y$ and $e_X = M_\perp X$, which is exactly the FWL estimator. □

4 Double Machine Learning (DML)

4.1 Motivation and Setup

Double Machine Learning (Chernozhukov et al., 2018) extends the FWL approach by:

1. Using flexible machine learning methods for nuisance function estimation
2. Employing cross-fitting to avoid overfitting bias
3. Providing valid inference despite high-dimensional nuisance parameters

4.2 DML Algorithm for Mediation

4.3 The DML Reduction Formula

A key contribution is the reduction formula that expresses PoMA in terms of prediction accuracies:

Algorithm 2 DML Mediation Analysis

Input: Data (X, M, Y), number of folds K
Step 1: Randomly partition data into K folds
for $k = 1$ to K **do**
  Train $\hat{g}_k : M \to Y$ on all folds except k
  Train $\hat{h}_k : M \to X$ on all folds except k
  Predict on fold k: $\hat{Y}_k = \hat{g}_k(M_k)$, $\hat{X}_k = \hat{h}_k(M_k)$
  Compute residuals: $e_{Y,k} = Y_k - \hat{Y}_k$, $e_{X,k} = X_k - \hat{X}_k$
end for
Step 2: Pool residuals across folds
Step 3: Estimate $c' = \text{Cov}(e_Y, e_X)/\text{Var}(e_X)$
Return: Direct effect c' with valid standard errors

Theorem 4 (DML Reduction Formula)**.** *The ratio of direct to total effect can be expressed as:*

$$\frac{c'}{c} = \frac{1 - \frac{Cov(\hat{Y},\hat{X})}{Cov(Y,X)} - C_1 - C_2}{1 - \frac{Var(\hat{X})}{Var(X)} - C_3} \tag{10}$$

where:

$$C_1 = \frac{Cov(e_Y, \hat{X})}{Cov(Y, X)} \tag{11}$$

$$C_2 = \frac{Cov(e_X, \hat{Y})}{Cov(Y, X)} \tag{12}$$

$$C_3 = \frac{2Cov(e_X, \hat{X})}{Var(X)} \tag{13}$$

Proof. Starting from the definition $c' = \text{Cov}(e_Y, e_X)/\text{Var}(e_X)$ and $c = \text{Cov}(Y, X)/\text{Var}(X)$:

$$\text{Cov}(e_Y, e_X) = \text{Cov}(Y - \hat{Y}, X - \hat{X}) \tag{14}$$

$$= \text{Cov}(Y, X) - \text{Cov}(Y, \hat{X}) - \text{Cov}(\hat{Y}, X) + \text{Cov}(\hat{Y}, \hat{X}) \tag{15}$$

Using the fact that $\text{Cov}(Y, \hat{X}) = \text{Cov}(\hat{Y}, X)$ under correct specification:

$$\text{Cov}(e_Y, e_X) = \text{Cov}(Y, X) - 2\text{Cov}(Y, \hat{X}) + \text{Cov}(\hat{Y}, \hat{X}) \tag{16}$$

Rearranging and noting that $\text{Cov}(Y, \hat{X}) = \text{Cov}(Y, X) - \text{Cov}(e_Y, \hat{X})$:

$$\text{Cov}(e_Y, e_X) = \text{Cov}(Y, X) \left[1 - \frac{\text{Cov}(\hat{Y}, \hat{X})}{\text{Cov}(Y, X)} - C_1 - C_2\right] \tag{17}$$

Similarly for the denominator:

$$\text{Var}(e_X) = \text{Var}(X)\left[1 - \frac{\text{Var}(\hat{X})}{\text{Var}(X)} - C_3\right] \tag{18}$$

Taking the ratio completes the proof. □

5 Causal Mediation Framework

5.1 Potential Outcomes and Natural Effects

The causal mediation framework (Pearl, 2001; Imai et al., 2010) uses potential outcomes notation:

- $Y(x, m)$: potential outcome under treatment x and mediator m
- $M(x)$: potential mediator value under treatment x
- $Y(x, M(x'))$: potential outcome under treatment x with mediator at its value under treatment x'

Definition 1 (Natural Direct Effect)**.** *For binary treatment:*

$$NDE = \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0, M(0))] \tag{19}$$

This is the effect of changing treatment from 0 to 1 while holding the mediator at its natural value under control.

Definition 2 (Natural Indirect Effect)**.**

$$NIE = \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(1, M(0))] \tag{20}$$

This is the effect of changing the mediator from its value under control to its value under treatment, while holding treatment at 1.

5.2 Key Properties

Proposition 5 (Effect Decomposition)**.** *The total effect decomposes as:*

$$\textit{Total Effect} = NDE + NIE \tag{21}$$

Proof.

$$\text{Total} = \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(0, M(0))] \tag{22}$$

$$= \underbrace{\mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(1, M(0))]}_{\text{NIE}} + \underbrace{\mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0, M(0))]}_{\text{NDE}} \tag{23}$$

□

5.3 Handling Interactions

A crucial advantage of natural effects is their ability to handle treatment-mediator interactions.

Example 1 (Linear Model with Interaction)**.** *Consider the model:*

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_3 XM + \epsilon \tag{24}$$

The controlled direct effect (CDE) at mediator level m is:

$$CDE(m) = \gamma_1 + \gamma_3 m \tag{25}$$

Note that CDE depends on m when $\gamma_3 \neq 0$. In contrast:

$$NDE = \gamma_1 + \gamma_3 \mathbb{E}[M(0)] \tag{26}$$
$$NIE = (\gamma_2 + \gamma_3)(\mathbb{E}[M(1)] - \mathbb{E}[M(0)]) \tag{27}$$

The natural effects provide a unique decomposition even with interaction.

6 Relationships Between Frameworks

6.1 When Methods Agree

Theorem 6 (Equivalence Conditions)**.** *The following are equivalent when:*

1. *All relationships are linear*
2. *No treatment-mediator interaction exists*
3. *Models are correctly specified*

Then: Traditional = FWL = DML (with linear models) = Natural Effects

Proof. Under linearity with no interaction:

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 M + \epsilon \tag{28}$$

All methods estimate γ_1 as the direct effect:

- Traditional: OLS coefficient of X controlling for M
- FWL: Coefficient from residual regression
- DML: Same as FWL but with cross-fitting (reduces to OLS under linearity)
- Natural Effects: NDE $= \gamma_1$ and CDE$(m) = \gamma_1$ for all m

□

6.2 When Methods Diverge

Proposition 7 (Non-linear Relationships)**.** *When the true relationship* $\mathbb{E}[Y|M]$ *is non-linear:*

- *Traditional and FWL give biased estimates*
- *DML with flexible ML remains consistent*
- *Natural effects require correct outcome model specification*

Proposition 8 (Interactions)**.** *When treatment-mediator interaction exists:*

- *Traditional, FWL, and DML estimate CDE at some weighted average of* M
- *Only natural effects provide meaningful decomposition*
- *CDE* $\neq$ *NDE in general*

7 The PoMA Instability Problem

7.1 Symmetric Relationships

Consider the symmetric non-linear relationship:

$$X \sim \text{Uniform}(-2, 2) \tag{29}$$

$$M = X^2 + \epsilon_M \tag{30}$$

$$Y = \sin(M) + \epsilon_Y \tag{31}$$

Proposition 9 (Near-Zero Covariance)**.** *In the above setup,* $Cov(X, Y) \approx 0$ *due to symmetry.*

Proof. By the law of total expectation:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{32}$$

$$= \mathbb{E}[X \sin(X^2)] - 0 \cdot \mathbb{E}[\sin(X^2)] \tag{33}$$

$$= \mathbb{E}[X \sin(X^2)] \tag{34}$$

Since $X \sin(X^2)$ is an odd function and X is symmetric around 0:

$$\mathbb{E}[X \sin(X^2)] = \int_{-2}^{2} x \sin(x^2) \frac{1}{4} dx = 0 \tag{35}$$

□

Corollary 10 (PoMA Explosion)**.** *When* $Cov(X, Y) \approx 0$*, the PoMA formula becomes:*

$$PoMA = 1 - \frac{c'}{c} = 1 - \frac{Cov(e_Y, e_X)/\,Var(e_X)}{Cov(Y, X)/\,Var(X)} \approx 1 - \frac{finite}{0} \tag{36}$$

This leads to extreme values like -35% *or* 1272%*.*

7.2 Other Edge Cases

Example 2 (Suppression Effect)**.** *When direct and indirect effects have opposite signs:*

$$M = 2X + \epsilon_M \tag{37}$$
$$Y = -X + 0.8M + \epsilon_Y \tag{38}$$

Here:

- *Direct effect:* -1
- *Indirect effect:* $2 \times 0.8 = 1.6$
- *Total effect:* 0.6
- *PoMA:* $1.6/0.6 = 267\%$

Remark 1. *PoMA* $> 100\%$ *is mathematically valid and indicates suppression, where the indirect path amplifies beyond the total effect due to opposing direct effects.*

8 Practical Recommendations

8.1 Diagnostic Procedure

Before conducting mediation analysis:

1. **Check correlations**: If $|\text{Cov}(X, Y)| < 0.05 \times \text{SD}(X) \times \text{SD}(Y)$, PoMA will be unstable
2. **Test linearity**: Compare R^2 of linear vs. polynomial models
3. **Test interactions**: Include XM term and check significance
4. **Check sample size**: ML methods need $n > 500$ typically

8.2 Method Selection Guide

8.3 Reporting Guidelines

1. Always report effect sizes, not just PoMA
2. Include confidence intervals
3. Check and report diagnostics
4. Use multiple methods for robustness
5. Be transparent about edge cases

Scenario	Recommended Method	Reason
Linear, no interaction	Traditional/FWL	Simple, proven
Non-linear relationships	DML with ML	Handles complexity
Treatment-mediator interaction	Natural Effects	Proper decomposition
Small sample ($n < 200$)	Traditional	Avoid overfitting
Near-zero effects	Any + Bootstrap CI	Quantify uncertainty

Table 1: Method selection guide

9 Conclusion

This theoretical framework unifies four major approaches to mediation analysis:

- Traditional and FWL are equivalent, both assuming linearity
- DML extends these to handle non-linearity via ML
- Natural effects provide the most general framework for interactions
- PoMA can be unstable in edge cases and should be interpreted carefully

The key insight is that no single method dominates in all scenarios. Practitioners should understand the assumptions and limitations of each approach and choose based on their specific context.

References

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

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

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420.

A Simulation Code Examples

A.1 Symmetric Relationship Demonstration

```
import numpy as np

# Generate symmetric data
np.random.seed(42)
n = 1000
X = np.random.uniform(-2, 2, n)
M = X**2 + np.random.normal(0, 0.2, n)
Y = np.sin(M) + np.random.normal(0, 0.1, n)

# Check correlation
print(f"Cor(X,Y) = {np.corrcoef(X, Y)[0,1]:.4f}")
# Output: Cor(X,Y) = 0.0156

# Traditional PoMA
from sklearn.linear_model import LinearRegression
lr1 = LinearRegression().fit(X.reshape(-1,1), Y)
c = lr1.coef_[0]
lr2 = LinearRegression().fit(np.column_stack([X, M]), Y)
c_prime = lr2.coef_[0]
poma = 1 - c_prime/c
print(f"PoMA = {poma:.1%}")
# Output: PoMA = -3540.0%
```

A.2 DML Implementation

```
from sklearn.model_selection import KFold
from sklearn.ensemble import RandomForestRegressor

def dml_mediation(X, M, Y, n_folds=5):
    kf = KFold(n_splits=n_folds, shuffle=True, random_state=42)

    # Cross-fitting
    Y_hat = np.zeros(len(Y))
    X_hat = np.zeros(len(X))

    for train_idx, test_idx in kf.split(X):
        # Train on folds != k
        rf_y = RandomForestRegressor(n_estimators=100)
        rf_x = RandomForestRegressor(n_estimators=100)

        rf_y.fit(M[train_idx].reshape(-1,1), Y[train_idx])
```

```
        rf_x.fit(M[train_idx].reshape(-1,1), X[train_idx])

        # Predict on fold k
        Y_hat[test_idx] = rf_y.predict(M[test_idx].reshape(-1,1))
        X_hat[test_idx] = rf_x.predict(M[test_idx].reshape(-1,1))

    # Compute residuals and direct effect
    e_Y = Y - Y_hat
    e_X = X - X_hat
    direct_effect = np.cov(e_Y, e_X)[0,1] / np.var(e_X)

    return direct_effect
```