Regression Imputation and Natural Mediation Effects

Oaxaca–Blinder, the mediation formula, and the three-regression decomposition

2026-11-16

Abstract Regression imputation provides a simple estimator of the mediation functional: fit the outcome regression among treated units and average its predictions over the control group's joint mediator distribution. Under sequential ignorability and overlap, subtracting the control outcome mean identifies the pure natural direct effect; the remaining portion of the total effect is the total natural indirect effect. This is the causal interpretation of one orientation of the Oaxaca–Blinder decomposition. The common-slope linear model gives the familiar Baron–Kenny product-of-coefficients formula. Separate outcome regressions accommodate treatment–mediator interactions and have an exact representation as an interacted pooled regression, whose coefficients yield both natural direct and indirect effects. Kline's double robustness concerns estimation of the observed-data functional. It does not replace the cross-world assumptions needed for natural effects. Diagonalizing mediator covariance preserves linear imputation and leaves this identification problem unchanged.

PDF · Executable Quarto source · Replication project · Numerical results

1 Regression transport as a mediation estimator

Consider a binary treatment D, an outcome Y, and a vector of intermediate variables $M \in \mathbb{R}^K$. Regression imputation learns how outcomes vary with M in one treatment arm and evaluates that relationship at mediator values observed in the other arm. The procedure combines an outcome response surface with a mediator distribution. This is precisely the structure of the **mediation formula** [1], [2].

Throughout, X denotes pre-treatment covariates, while $W = (1, M')'$ denotes the intercept-augmented mediator regressors.

The direction of the imputation determines the natural effect. Fitting the treated outcome regression and evaluating it on control mediators estimates the nested mean $\mathbb{E}[Y(1, M(0))]$, provided the identifying assumptions below hold. Subtracting the control outcome mean gives the **pure natural direct effect**. Subtracting the imputed mean from the treated outcome mean gives the **total natural indirect effect**.

Three different statements are involved:

1. The fitted regression induces an exact sample decomposition.
2. The regression estimates a particular observed-data functional.
3. The functional identifies a causal natural effect under additional assumptions.

The first statement is algebraic. The second involves statistical specification and support. The third involves mediator–outcome confounding and cross-world counterfactuals. A small, well-conditioned mediator vector helps estimation; it does not by itself establish the third statement.

The main derivation uses a randomized treatment without baseline covariates, so arm-specific empirical averages target a common population. Section 7 gives the conditional version with pre-treatment covariates X. All effects are defined on the **mean-difference scale**.

2 Potential outcomes and the mediation formula

2.1 Nested means and the two decompositions

Write $M(d)$ for the entire mediator vector under treatment d, and $Y(d, m)$ for the outcome under a joint intervention setting treatment to d and mediators to m. Consistency and composition give

$$M = M(D), \qquad Y = Y(D, M(D)).$$

Define

$$\Psi_{dd'} = \mathbb{E}[Y(d, M(d'))], \qquad d, d' \in \{0, 1\}. \tag{1}$$

The total effect is $\Psi_{11} - \Psi_{00}$. Adding and subtracting either cross-world mean gives two decompositions:

$$\begin{aligned} \mathrm{TE} &= \underbrace{\Psi_{10} - \Psi_{00}}_{\mathrm{NDE}_0} + \underbrace{\Psi_{11} - \Psi_{10}}_{\mathrm{NIE}_1}, \\ &= \underbrace{\Psi_{11} - \Psi_{01}}_{\mathrm{NDE}_1} + \underbrace{\Psi_{01} - \Psi_{00}}_{\mathrm{NIE}_0}. \end{aligned} \tag{2}$$

Effect	Counterfactual contrast	Conventional name
NDE_0	$\Psi_{10} - \Psi_{00}$	Pure natural direct effect
NIE_1	$\Psi_{11} - \Psi_{10}$	Total natural indirect effect
NDE_1	$\Psi_{11} - \Psi_{01}$	Total natural direct effect
NIE_0	$\Psi_{01} - \Psi_{00}$	Pure natural indirect effect

The subscripts indicate the mediator reference arm for direct effects and the fixed outcome-treatment arm for indirect effects. With treatment–mediator interactions, the two direct effects and the two indirect effects generally differ.

2.2 Sufficient identifying assumptions

Let

$$h_d(m, x) = \mathbb{E}[Y \mid D = d, M = m, X = x], \qquad g_d(m \mid x) = f_{M \mid D=d, X=x}(m).$$

One transparent, strong set of sufficient conditions is:

1. **Consistency, composition, and well-defined joint interventions.** The observed mediator and outcome equal their corresponding potential values, with no interference between units.
2. **Treatment exchangeability.** $\{Y(d, m), M(d')\} \perp D \mid X$ for all relevant d, d', m. Random assignment supplies this condition when X is omitted.
3. **Mediator–outcome and cross-world exchangeability.** $Y(d, m) \perp M(d') \mid X$ for all relevant d, d', m. This is a sufficient version of the cross-world requirement; mean independence can suffice for the mean targets here.
4. **Positivity.** Both treatment arms occur in the relevant baseline strata, and $g_1(m \mid x) > 0$ wherever the control mediator distribution contributes to Ψ_{10}. Reverse overlap is needed for Ψ_{01}.

These conditions imply the usual sequential-ignorability formulation used for mediation [1], [2]. In a structural model with independent mediator and outcome disturbances conditional on X, the cross-world condition follows from the structural assumptions. An unmeasured common cause of M and Y, or an unaccounted treatment-induced mediator–outcome confounder, invalidates this route. **Randomizing D alone does not randomize M.**

The identifying calculation is

$$\begin{aligned} \mathbb{E}[Y(1, M(0)) \mid X = x] &= \int \mathbb{E}[Y(1, m) \mid M(0) = m, X = x]\, g_0(m \mid x)\, dm \\ &= \int \mathbb{E}[Y(1, m) \mid X = x]\, g_0(m \mid x)\, dm \\ &= \int h_1(m, x)\, g_0(m \mid x)\, dm. \end{aligned} \tag{3}$$

Consequently,

$$\boxed{\Psi_{10} = \mathbb{E}_X\left[\int h_1(m, X)\, g_0(m \mid X)\, dm\right].} \tag{4}$$

The middle equality in Equation 3 uses the cross-world condition. The last equality uses exchangeability and consistency to recover the interventional response from the observed regression. Integrating over an observed mediator distribution establishes neither equality on its own.

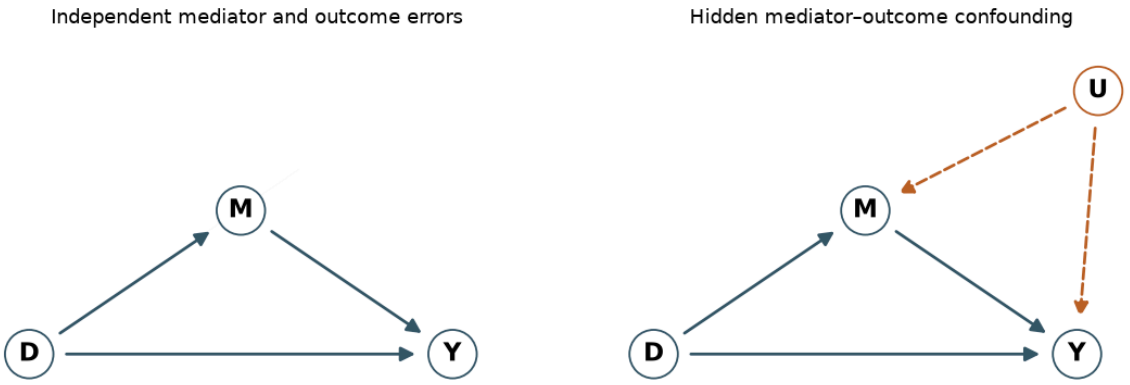


Figure 1: Randomized treatment and mediator identification. The left graph admits the usual mediation formula under independent disturbances and consistency. The right graph adds an unmeasured common cause of the joint mediator vector and the outcome; randomization of D remains intact, but the regression no longer recovers an intervention on M. Independence between components of M does not remove the U paths.

3 The Oaxaca–Blinder estimator

3.1 Empirical integration over control mediators

Suppress X and let $W = (1, M')'$. Fit the treated regression

$$\hat{\beta}_1 = \arg\min_{\beta} \sum_{i:D_i=1} (Y_i - W_i'\beta)^2, \qquad \hat{h}_1(m) = \hat{\alpha}_1 + m'\hat{b}_1.$$

Replacing the control mediator distribution by its empirical distribution gives

$$\hat{\Psi}_{10}^{\mathrm{OB}} = \frac{1}{n_0} \sum_{i:D_i=0} \hat{h}_1(M_i) = \hat{\alpha}_1 + \overline{M}_0'\hat{b}_1. \tag{5}$$

Thus, under identification and suitable statistical specification,

$$\begin{aligned} \widehat{\mathrm{NDE}}_0 &= \hat{\Psi}_{10}^{\mathrm{OB}} - \overline{Y}_0, \\ \widehat{\mathrm{NIE}}_1 &= \overline{Y}_1 - \hat{\Psi}_{10}^{\mathrm{OB}}. \end{aligned} \tag{6}$$

The imputed mean is a nested outcome level. It becomes a direct effect only after subtracting the control mean. With post-treatment mediators and these assumptions, the interpretation is a natural direct effect. The ATU interpretation used for imputation on pre-treatment adjustment variables does not carry over unchanged.

3.2 Exact sample decomposition

Fit a control regression as well. Because both regressions contain intercepts, $\overline{Y}_d = \hat{\alpha}_d + \overline{M}_d'\hat{b}_d$. The sample difference in means then decomposes exactly:

$$\begin{aligned} \overline{Y}_1 - \overline{Y}_0 = {} & \underbrace{(\hat{\alpha}_1 - \hat{\alpha}_0) + \overline{M}_0'\big(\hat{b}_1 - \hat{b}_0\big)}_{\text{direct-effect estimate under identification}} \\ & + \underbrace{\big(\overline{M}_1 - \overline{M}_0\big)'\hat{b}_1}_{\text{indirect-effect estimate under identification}}. \end{aligned} \tag{7}$$

This is the familiar twofold Oaxaca–Blinder decomposition using treated outcome coefficients as the reference. The first component is often called the *unexplained* component; the second is the *explained* or composition component. Their natural-effect labels require the assumptions above. The control regression is useful for writing the decomposition, but it is not needed to compute Equation 6.

Reversing the outcome reference arm gives $\hat{\Psi}_{01} = \hat{\alpha}_0 + \overline{M}_1'\hat{b}_0$ and the alternative pair $\left(\widehat{\mathrm{NDE}}_1, \widehat{\mathrm{NIE}}_0\right)$.

A nonlinear fitted response surface also admits empirical integration: $n_0^{-1}\sum_{D_i=0}\hat{h}_1(M_i)$. The mediation formula survives, although the mean-vector simplification and the specific OLS reweighting algebra below may no longer apply.

4 Comparison with the three Baron–Kenny regressions

4.1 The common-slope linear model

The familiar three-regression setup [3], extended to a vector mediator, is

$$\begin{aligned} Y_i &= a_Y + cD_i + e_i, && \text{total-outcome regression,} \\ M_i &= \mu + aD_i + u_i, && \text{mediator regressions,} \\ Y_i &= \alpha + c'D_i + b'M_i + v_i, && \text{adjusted-outcome regression .} \end{aligned} \tag{8}$$

The second line represents K equations with the same right-hand side. Under the linear causal model, the conditional mean restrictions needed for these equations, and mediation identification,

$$\mathrm{NDE}_0 = \mathrm{NDE}_1 = c', \qquad \mathrm{NIE}_0 = \mathrm{NIE}_1 = a'b, \qquad c = c' + a'b. \tag{9}$$

The OLS identity $\hat{c} = \hat{c}' + \hat{a}'\hat{b}$ holds exactly when the three regressions use the same observations, intercepts, and no differing weights. This identity follows from linear projection, so it holds even when the causal interpretation fails.

In the correctly specified common-slope model, treated-arm regression imputation and the Baron–Kenny coefficients target the **same natural effects**. Separate-arm OLS and restricted pooled OLS generally give different finite-sample estimates because they impose different slope restrictions.

The historical causal-steps testing procedure is a separate matter. Evidence for an indirect effect does not require a statistically significant total effect: direct and indirect effects can offset each other. Estimating and obtaining uncertainty for the relevant natural-effect contrast is more informative than requiring each coefficient test to pass in sequence [1].

4.2 Treatment–mediator interactions

Allow

$$Y(d,m) = \alpha + \tau d + b'm + d\,\kappa'm + \varepsilon, \qquad M(d) = \mu + ad + u, \tag{10}$$

with mean-zero disturbances satisfying the identifying restrictions. Now the controlled direct effect at m is $\tau + \kappa'm$, and the natural effects are

$$\begin{aligned} \mathrm{NDE}_0 &= \tau + \kappa'\mu, && \mathrm{NIE}_1 = (b+\kappa)'a, \\ \mathrm{NDE}_1 &= \tau + \kappa'(\mu + a), && \mathrm{NIE}_0 = b'a. \end{aligned} \tag{11}$$

Separate treated and control regressions permit different slopes $b_1 = b + \kappa$ and $b_0 = b$. They therefore recover the correct natural-effect pair in this model. A Baron–Kenny third regression that omits $D \times M$ imposes a false common slope.

For a particularly clear case, suppose $P(D=1) = p$ and $\mathrm{Var}\,(M \mid D=0) = \mathrm{Var}\,(M \mid D=1) = \Sigma$. The common-slope pooled regression then has population coefficients

$$b_{\mathrm{pool}} = b + p\kappa, \qquad c'_{\mathrm{pool}} = \mathrm{NDE}_0 + (1-p)\kappa'a. \tag{12}$$

Even when Σ is diagonal, its treatment coefficient generally differs from the pure natural direct effect.

4.3 An exact interacted-regression representation

There is a useful finite-sample equivalence. Define $Z_i = M_i - \overline{M}_0$ and fit the fully interacted pooled regression

$$Y_i = A + \delta D_i + B' Z_i + D_i H' Z_i + e_i. \tag{13}$$

This is a reparameterization of the two separate outcome regressions. It uses the fully interacted form familiar from W. Lin [4], but centers post-treatment mediators at the control mean. Lin's randomization-based results concern adjustment for pre-treatment covariates; the mediation interpretation here additionally requires the identification conditions in Equation 3. At the control mediator mean, the fitted arm difference is

$$\boxed{\hat{\delta} = \hat{\Psi}_{10}^{\mathrm{OB}} - \overline{Y}_0 = \widehat{\mathrm{NDE}}_0.} \tag{14}$$

Thus the direct-effect estimate can be read directly from a treatment coefficient, provided the mediators are centered at the control mean and all treatment–mediator interactions are included. The equality is algebraic; its causal interpretation still needs identification.

4.4 Natural indirect effects from the same fit

The interacted regression also supplies both natural indirect effects. Let $\Delta_M = \overline{M}_1 - \overline{M}_0$. The fitted outcome slopes are $\hat{B}$ in controls and $\hat{B} + \hat{H}$ in treated units. The four estimates are

$$\begin{aligned} \widehat{\mathrm{NDE}}_0 &= \hat{\delta}, & \widehat{\mathrm{NIE}}_1 &= \left(\hat{B} + \hat{H}\right)' \Delta_M, \\ \widehat{\mathrm{NDE}}_1 &= \hat{\delta} + \hat{H}' \Delta_M, & \widehat{\mathrm{NIE}}_0 &= \hat{B}' \Delta_M. \end{aligned} \tag{15}$$

The first row uses the control mediator distribution to define the direct effect and treated outcome slopes to define the indirect effect. The second row gives the complementary pair. Both come from the same coefficient vector and observed mediator means.

To verify the decomposition, the OLS normal equations within the two treatment arms give

$$\begin{aligned} \overline{Y}_0 &= \hat{A}, \\ \overline{Y}_1 &= \hat{A} + \hat{\delta} + \left(\hat{B} + \hat{H}\right)' \Delta_M. \end{aligned}$$

Consequently,

$$\begin{aligned} \overline{Y}_1 - \overline{Y}_0 &= \hat{\delta} + \left(\hat{B} + \hat{H}\right)' \Delta_M \\ &= \left(\hat{\delta} + \hat{H}' \Delta_M\right) + \hat{B}' \Delta_M. \end{aligned} \tag{16}$$

Thus both natural-effect pairs exactly decompose the sample outcome gap. The allocation across the two reference choices differs by $\hat{H}' \Delta_M$. Imposing $H = 0$ recovers the common-slope Baron–Kenny decomposition.

With binary D and no baseline covariates, a separate regression of M on an intercept and D would give treatment coefficient $\hat{a} = \Delta_M$. The indirect components are therefore the corresponding products of the mediator shift and outcome slope; no additional mediator fit is required in this setup. With baseline X, population-average effects require the common-population standardization described in Section 7.

For standard errors, both the mediator shift and the control centering mean are estimated. A pairs bootstrap that recomputes these quantities, refits the interacted regression, and recalculates all four contrasts captures their joint uncertainty. The regression's usual treatment coefficient standard error does not by itself give uncertainty for the indirect effects.

Note 1: Special case: equal natural direct and indirect effects

Suppose the pure and total natural direct effects coincide, and likewise for the indirect effects. These are **one restriction**, because the potential-outcome definitions imply

$$\mathrm{NDE}_1 - \mathrm{NDE}_0 = \mathrm{NIE}_1 - \mathrm{NIE}_0. \tag{17}$$

Both differences measure the average treatment–mediator interaction on the additive outcome scale. Setting either difference to zero sets the other to zero, leaving a single natural direct effect and a single natural indirect effect.

For the population version of Equation 13, center at $\mathbb{E}[M(0)]$ and write $\Delta = \mathbb{E}[M(1)] - \mathbb{E}[M(0)]$. In this linear model the restriction is $H'\Delta = 0$, so

$$\begin{aligned} \mathrm{NDE}_0 = \mathrm{NDE}_1 &= \delta, \\ \mathrm{NIE}_0 = \mathrm{NIE}_1 &= B'\Delta = (B+H)'\Delta, \\ \mathrm{TE} &= \delta + B'\Delta. \end{aligned} \tag{18}$$

Two versions of the simplifying assumption should be distinguished:

- **No interactions:** $H = 0$ eliminates all treatment–mediator interactions in the linear outcome model and gives the common-slope Baron–Kenny specification and its product-of-coefficients indirect effect.
- **Zero average interaction:** only $H'\Delta = 0$ is imposed. Interactions can remain but cancel along the average mediator shift. For example, $H = (1, -1)'$ and $\Delta = (1, 1)'$ satisfy this restriction. It is a single scalar condition, rather than the vector restriction $H = 0$. A pooled common-slope regression is not generally justified by this weaker condition; the interacted model can be retained.

The equality concerns population effects. Unrestricted sample estimates can differ through sampling variation. The mediation identification conditions in Equation 3 remain necessary for their causal interpretation, including control of mediator–outcome confounding.

5 Mediator covariance and causal identification

5.1 Diagonal covariance simplifies a projection

Let $\Sigma_1 = \mathrm{Var}\ (M \mid D = 1)$. The population treated projection slope is

$$b_1^* = \Sigma_1^{-1}\ \mathrm{Cov}\ (M, Y \mid D = 1).$$

If Σ_1 is diagonal, each coefficient reduces to

$$b_{1j}^* = \frac{\mathrm{Cov}\ \left(M_j, Y \mid D = 1\right)}{\mathrm{Var}\ \left(M_j \mid D = 1\right)}.$$

This is a statement about correlations among the mediators and the conditioning of the regression. It places no restriction on unmeasured causes shared by a mediator and the outcome. Even mutual independence of mediator components would leave that problem open.

5.2 A diagonal-covariance counterexample

Let D be randomized, let U and E be independent standard Gaussian vectors, and let the mean-zero outcome disturbance η be independent of D, U, E. In particular, $\mathrm{Var}\ (U) = \mathrm{Var}\ (E) = I$. Consider

$$\begin{aligned} M &= \mu + aD + \frac{U+E}{\sqrt{2}}, \\ Y &= \alpha + \tau D + b'M + \gamma'U + \eta. \end{aligned} \tag{19}$$

The mediator covariance in **each treatment arm is exactly I**. The structural natural direct effect is τ, because the direct treatment effect is constant. Nevertheless,

$$b_1^* = b + \frac{\gamma}{\sqrt{2}}, \qquad \operatorname{plim} \widehat{\mathrm{NDE}}_0^{\mathrm{OB}} = \tau - \frac{a'\gamma}{\sqrt{2}}. \tag{20}$$

The regression attributes part of the U–outcome association to M. Transporting its slope across treatment arms transfers that error into the direct/indirect split. The total difference in means remains a valid randomized-treatment effect.

5.3 Whitening preserves the imputation

For any nonsingular A and vector c, replacing M by $\widetilde{M} = A(M - c)$ leaves the intercept-plus-linear regressor space unchanged. OLS fitted values, transported predictions, and both Oaxaca–Blinder components are identical.

In the counterexample,

$$\mathrm{Var}\,(M) = I + p(1-p)aa'.$$

Whitening by this pooled covariance makes $\mathrm{Var}\left(\widetilde{M}\right) = I$ **unconditionally**, while leaving the biased estimate unchanged. A sample whitening transformation has the same exact fitted-value invariance. This directly covers imposing diagonality on the pooled covariance rather than the arm-specific one. Dropping principal components or imposing a new ridge penalty changes the estimator and requires a separate approximation analysis.

6 Kline's robustness property for the mediation functional

6.1 Regression imputation as balancing weights

With $W_i = (1, M_i')'$ and treated design matrix W_1, Equation 5 can be rewritten

$$\hat{\Psi}_{10}^{\mathrm{OB}} = \sum_{i:D_i=1} w_i Y_i, \qquad w_i = W_i'(W_1'W_1)^{-1}\overline{W}_0. \tag{21}$$

These weights satisfy

$$\sum_{D_i=1} w_i = 1, \qquad \sum_{D_i=1} w_i W_i = \overline{W}_0.$$

They exactly balance the included mediator basis. Some weights can be negative. The algebra is the same regression/reweighting equivalence studied by P. Kline [5], now applied to distributions of intermediate variables.

Define the **observed-data functional**

$$\Theta = \int h_1(m)\, g_0(m)\, dm$$

and the density ratio $r(m) = g_0(m)/g_1(m)$. Under overlap, fixed-dimensional OLS imputation is consistent for Θ if either:

- $h_1(m)$ lies in the fitted linear span; or
- $r(m)$ lies in that linear span, even if $h_1(m)$ does not.

The second branch is a linear model for a density ratio. By Bayes' rule,

$$r(m) = \frac{p}{1-p}\,\frac{1-\pi_M(m)}{\pi_M(m)}, \qquad \pi_M(m) = P(D=1 \mid M=m). \tag{22}$$

Here π_M is a **post-treatment classification probability**. It differs from the randomized assignment probability p, or the baseline propensity score $P(D=1 \mid X)$. Random assignment does not make π_M constant when treatment changes the mediators. A logistic model with a linear index makes odds exponential in the regressors; it generally does not satisfy this linear-odds condition.

6.2 The product of approximation errors

Let $\tilde{h}_1$ denote the population linear projection of h_1 under the treated mediator distribution, and $\tilde{r}$ the corresponding linear projection of r. The population imputation bias is

$$\tilde{\Theta} - \Theta = \mathbb{E}\big[(\tilde{r}(M) - r(M))\big(h_1(M) - \tilde{h}_1(M)\big)\big], \tag{23}$$

where $\mathbb{E}_1$ averages over $M \mid D = 1$. To see this, write the bias as $-\mathbb{E}_1\left[r\left(h_1 - \tilde{h}_1\right)\right]$ and use the projection orthogonality of $h_1 - \tilde{h}_1$ to every included regressor, including $\tilde{r}$. Either exact specification makes the product zero. Near-correct specification can also help when the two residuals have a small inner product.

This robustness concerns estimation of Θ. Its equality to Ψ_{10} still requires Equation 3. In the confounded counterexample, h_1 is exactly linear: OLS consistently estimates the observed-data functional, but that functional gives the wrong causal direct effect.

6.3 Relation to augmented mediation estimators

An augmented estimator of the same no-covariate functional is

$$\hat{\Theta}_{\text{aug}} = \frac{1}{n_0} \sum_{D_i=0} \hat{h}_1(M_i) + \frac{1}{n_1} \sum_{D_i=1} \hat{r}(M_i)\left\{Y_i - \hat{h}_1(M_i)\right\}. \tag{24}$$

With suitable regularity, it is consistent if either the outcome regression or the density ratio is consistently estimated. For OLS $\hat{h}_1$ and $\hat{r}(M_i) = n_1 w_i$, the augmentation is **exactly zero** by the OLS normal equations. This exposes the connection between ordinary regression imputation and augmented estimators.

The conditional-covariate setting adds nuisance functions for treatment assignment and mediator distributions. E. J. Tchetgen Tchetgen and I. Shpitser [2] develop multiply robust, locally efficient estimators for that general mediation functional. Those estimators offer statistical robustness under their nuisance-model conditions; they retain the causal identifying assumptions. Ordinary OB imputation is not automatically semiparametrically efficient.

7 Baseline covariates, multiple mediators, and measurements

7.1 Standardizing to a common baseline population

With baseline confounders X, the target is Equation 4. A regression on (M, X) followed by an unweighted average over controls targets the **control baseline distribution**. If D is randomized, that distribution matches the population baseline distribution asymptotically. In an observational design, it generally does not.

For a population-average natural effect, estimate

$$\eta_{10}(x) = \int h_1(m, x) g_0(m \mid x)\, dm, \qquad \Psi_{10} = \mathbb{E}_X[\eta_{10}(X)].$$

One implementation estimates the mediator model in controls, integrates or simulates predictions conditional on each target unit's X_i, and averages over the target baseline sample. Another uses appropriate control inverse-propensity weights. For example, with $e(X) = P(D = 1 \mid X)$,

$$\Psi_{10} = \mathbb{E}\left[\frac{1 - D}{1 - e(X)} h_1(M, X)\right].$$

The corresponding Ψ_{00} must be standardized to the same baseline population. Subtracting the raw control outcome mean while standardizing only Ψ_{10} would mix populations.

7.2 Joint and component-specific mediation

For the joint vector M, empirical integration uses each control unit's **entire mediator vector**. It preserves dependence between components. No diagonal covariance assumption is needed.

This identifies a joint natural effect under the stated conditions. Attributing the joint indirect effect to individual components generally requires additional structure: causal ordering, mediator–mediator interactions, and counterfactual combinations such as $(M_1(0), M_2(1))$ matter. Diagonal within-arm covariance does not specify those cross-world combinations.

If individual-level cross-world counterfactuals are unsuitable, one can define a stochastic mediator draw G_0 from $g_0(\cdot \mid X)$, independent of a unit's potential outcomes given X. Under appropriate within-world exchangeability, the same functional identifies $\mathbb{E}[Y(1, G_0)]$. This gives an **interventional direct-effect analogue**, with a different intervention and interpretation. It still requires mediator–outcome confounding control; treatment-induced confounders require a more elaborate longitudinal g-formula.

7.3 A measurement can be a proxy for a mediator

A benchmark score or offline evaluation can be a measurement of an intermediate latent state rather than a causal mediator in its own right. For example,

$$L = aD + U, \qquad M = L + \epsilon, \qquad Y = \tau D + bL + \eta$$

has the paths $D \to L \to Y$ and $L \to M$. Changing a reported value of M while keeping L fixed does not change Y. Under that intervention, the indirect effect through the measurement M is zero, even though regression on M can absorb much of the treatment–outcome association.

With independent disturbances, regression imputation in this scalar model has probability limit

$$\mathrm{NDE}^{\mathrm{OB}} = \tau + ab\,\frac{\mathrm{Var}\,(\epsilon)}{\mathrm{Var}\,(U) + \mathrm{Var}\,(\epsilon)},$$

whereas the natural direct effect **through the observed readout** is the entire total effect $\tau + ab$. A mediation effect through the latent state L is a different estimand. Applying measurement or surrogate-index methods to learn about that state requires an explicit measurement/bridge model. Regression predictiveness alone does not choose among these targets.

8 Numerical illustrations

The following calculations reproduce the population algebra and illustrate the sampling behavior. All data are simulated for this note. Each of three designs has $K = 3$ mediators and randomized treatment. There are 300 independent samples per design, each with 4,000 observations. Reported Monte Carlo means average over those 300 independent samples.

8.1 Designs and estimators

The common-slope design sets $\kappa = \gamma = 0$. The interaction design sets $\kappa = (0.7, -0.4, 0.5)'$ and $\gamma = 0$. The confounded design sets $\kappa = 0$ and $\gamma = (0.8, -0.5, 0.6)'$. All use $\tau = 0.6$, $\mu = (0.8, -0.4, 0.5)'$, $a = (0.7, -0.5, 0.35)'$, and $b = (0.6, -0.4, 0.3)'$.

```
mu = np.array([0.8, -0.4, 0.5])
a = np.array([0.7, -0.5, 0.35])
b = np.array([0.6, -0.4, 0.3])
tau = 0.6
designs = {
    "Common slope": (np.zeros(3), np.zeros(3)),
    "Treatment interaction": (np.array([0.7, -0.4, 0.5]), np.zeros(3)),
    "Hidden confounding": (np.zeros(3), np.array([0.8, -0.5, 0.6])),
}

def simulate(n, kappa, gamma, rng):
    d = rng.binomial(1, 0.5, n)
    u = rng.normal(size=(n, 3))
    e = rng.normal(size=(n, 3))
    m = mu + d[:, None] * a + (u + e) / np.sqrt(2)
    y = (0.4 + tau * d + m @ b + d * (m @ kappa)
         + u @ gamma + rng.normal(size=n))
    return d, m, y

def ob_fit(d, m, y):
    w = np.column_stack([np.ones(len(d)), m])
    coef = np.linalg.lstsq(w[d == 1], y[d == 1], rcond=None)[0]
    nested = np.mean(w[d == 0] @ coef)
    direct = nested - y[d == 0].mean()
    indirect = y[d == 1].mean() - nested
    np.testing.assert_allclose(
        direct + indirect, y[d == 1].mean() - y[d == 0].mean(), atol=1e-10
    )
    return np.array([direct, indirect, direct + indirect])
```

```
def bk_fit(d, m, y):
    design_d = np.column_stack([np.ones(len(d)), d])
    total = np.linalg.lstsq(design_d, y, rcond=None)[0][1]
    mediator_shift = np.linalg.lstsq(design_d, m, rcond=None)[0][1]
    adjusted = np.column_stack([design_d, m])
    coef = np.linalg.lstsq(adjusted, y, rcond=None)[0]
    direct = coef[1]
    indirect = mediator_shift @ coef[2:]
    np.testing.assert_allclose(direct + indirect, total, atol=1e-10)
    return np.array([direct, indirect, total])

def natural_effects(kappa):
    direct = tau + mu @ kappa
    indirect = a @ (b + kappa)
    return np.array([direct, indirect, direct + indirect])
```

Both methods exactly decompose their sample total effect. The common-slope method imposes the restriction in the third Baron–Kenny regression. The OB method fits the treated response surface without borrowing the control outcome slope.

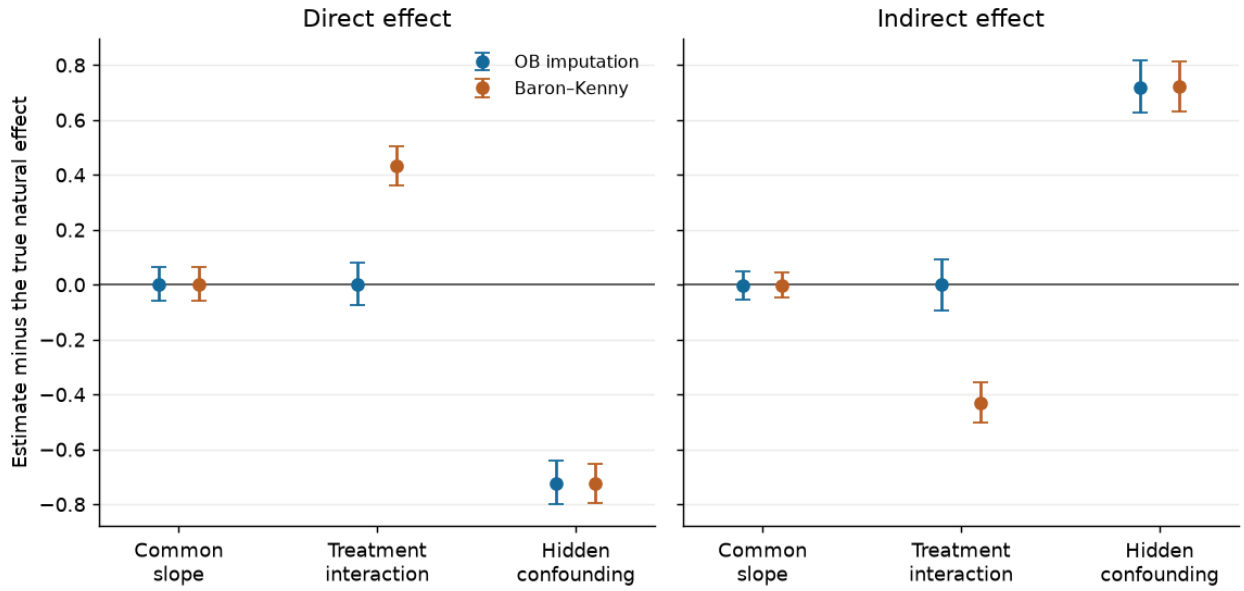


Figure 2: Direct- and indirect-effect estimation errors. Points are means across 300 independent samples; bars show the central 90% of the sampling distribution of the estimator, not a confidence interval for the Monte Carlo mean. Each sample has 4,000 observations and three mediators. The confounded design has diagonal within-arm population mediator covariance.

Table 1

Design	Estimator	Direct bias	Indirect bias	Total bias
Common slope	Baron–Kenny	+0.002	−0.002	+0.000
Common slope	OB imputation	+0.002	−0.002	+0.000
Hidden confounding	Baron–Kenny	−0.723	+0.721	−0.003
Hidden confounding	OB imputation	−0.723	+0.720	−0.003
Treatment interaction	Baron–Kenny	+0.434	−0.432	+0.002
Treatment interaction	OB imputation	+0.002	−0.000	+0.002

The interaction design has true direct effect 1.570 and indirect effect 1.590. The common-slope Baron–Kenny direct coefficient converges to 2.0025, overshooting the pure natural direct effect by 0.4325; the indirect estimate

undershoots by the same amount. The confounded design has true direct effect 0.600, but both procedures converge to $0.600 - 1.020/\sqrt{2} \simeq -0.121$. Their agreement does not establish valid mediation identification.

8.2 Exact equivalences and whitening checks

The next calculation verifies the centered interacted regression, exact covariate balance, and invariance to whitening the **pooled sample** mediator covariance. It uses one additional sample from the confounded design.

```
d, m, y = simulate(6000, *designs["Hidden confounding"], rng)
z = m - m[d == 0].mean(axis=0)
interacted = np.column_stack([np.ones(len(d)), d, z, d[:, None] * z])
delta = np.linalg.lstsq(interacted, y, rcond=None)[0][1]
ob_original = ob_fit(d, m, y)
np.testing.assert_allclose(delta, ob_original[0], atol=1e-10)

cov = np.cov(m, rowvar=False, ddof=0)
eigenvalues, eigenvectors = np.linalg.eigh(cov)
whitener = (eigenvectors * (1 / np.sqrt(eigenvalues))) @ eigenvectors.T
m_white = (m - m.mean(axis=0)) @ whitener
np.testing.assert_allclose(np.cov(m_white, rowvar=False, ddof=0),
                           np.eye(3), atol=1e-10)
np.testing.assert_allclose(ob_fit(d, m_white, y), ob_original, atol=1e-10)

w = np.column_stack([np.ones(len(d)), m])
w1, y1 = w[d == 1], y[d == 1]
wbar0 = w[d == 0].mean(axis=0)
weights = w1 @ np.linalg.solve(w1.T @ w1, wbar0)
coef = np.linalg.lstsq(w1, y1, rcond=None)[0]
coef0 = np.linalg.lstsq(w[d == 0], y[d == 0], rcond=None)[0]
twofold_direct = (coef[0] - coef0[0]
                  + m[d == 0].mean(axis=0) @ (coef[1:] - coef0[1:]))
np.testing.assert_allclose(twofold_direct, ob_original[0], atol=1e-10)
np.testing.assert_allclose(weights.sum(), 1, atol=1e-10)
np.testing.assert_allclose(weights @ w1, wbar0, atol=1e-10)
np.testing.assert_allclose(weights @ y1, wbar0 @ coef, atol=1e-10)
augmentation = np.mean(len(y1) * weights * (y1 - w1 @ coef))
assert abs(augmentation) < 1e-10

algebra = {
    "centered_coefficient_gap": float(abs(delta - ob_original[0])),
    "whitening_fit_gap": float(np.max(abs(ob_fit(d, m_white, y) - ob_original))),
    "whitening_covariance_gap": float(np.max(abs(
        np.cov(m_white, rowvar=False, ddof=0) - np.eye(3)))),
    "balance_gap": float(np.max(abs(weights @ w1 - wbar0))),
    "twofold_direct_gap": float(abs(twofold_direct - ob_original[0])),
    "augmentation": float(augmentation),
}
display(Markdown(
    f"The centered interacted coefficient and OB direct component differ by "
    f"{algebra['centered_coefficient_gap']:.2g}; pooled whitening changes the "
    f"estimated effects by at most {algebra['whitening_fit_gap']:.2g}. "
    f"The largest sample covariance deviation from the identity after "
    f"whitening is {algebra['whitening_covariance_gap']:.2g}."
))
```

The centered interacted coefficient and OB direct component differ by 2.9e-15; pooled whitening changes the estimated effects by at most 6.7e-16. The largest sample covariance deviation from the identity after whitening is 1.4e-15.

8.3 Checking both natural-effect decompositions

This calculation obtains all four effects from one interacted regression and compares them with both orientations of regression imputation. It checks each of the three designs, including the interaction case, using a separate random-number stream so the original Monte Carlo and bootstrap samples remain unchanged.

```
def lin_effects(d, m, y):
    k = m.shape[1]
    z = m - m[d == 0].mean(axis=0)
    design = np.column_stack([np.ones(len(d)), d, z, d[:, None] * z])
    coef = np.linalg.lstsq(design, y, rcond=None)[0]
    delta, control_slope, interaction = coef[1], coef[2:2+k], coef[2+k:]
    shift = m[d == 1].mean(axis=0) - m[d == 0].mean(axis=0)
    return np.array([
        delta,
        (control_slope + interaction) @ shift,
        delta + interaction @ shift,
        control_slope @ shift,
    ])

identity_rng = np.random.default_rng(20261117)
identity_rows = []
for case, (kappa, gamma) in designs.items():
    di, mi, yi = simulate(6000, kappa, gamma, identity_rng)
    effects = lin_effects(di, mi, yi)
    ob_forward = ob_fit(di, mi, yi)
    wi = np.column_stack([np.ones(len(di)), mi])
    beta0 = np.linalg.lstsq(wi[di == 0], yi[di == 0], rcond=None)[0]
    psi01 = np.mean(wi[di == 1] @ beta0)
    ob_reverse = np.array([yi[di == 1].mean() - psi01,
                           psi01 - yi[di == 0].mean()])
    imputation = np.r_[ob_forward[:2], ob_reverse]
    np.testing.assert_allclose(effects, imputation, atol=1e-10)
    total = yi[di == 1].mean() - yi[di == 0].mean()
    np.testing.assert_allclose([effects[:2].sum(), effects[2:].sum()],
                               [total, total], atol=1e-10)
    np.testing.assert_allclose(effects[2] - effects[0],
                               effects[1] - effects[3], atol=1e-10)
    for effect, estimate, reference in zip(
        ["NDE0", "NIE1", "NDE1", "NIE0"], effects, imputation
    ):
        identity_rows.append({
            "design": case, "effect": effect,
            "interacted_regression": estimate, "regression_imputation": reference,
            "absolute_gap": abs(estimate - reference),
        })
identity_results = pd.DataFrame(identity_rows)
identity_results.to_csv(RESULTS / "lin-effect-identities.csv", index=False)
identity_max_gap = float(identity_results.absolute_gap.max())
display(Markdown(
    f"All four contrasts agree with regression imputation in all three designs; "
    f"the largest absolute gap is {identity_max_gap:.2g}. Both decompositions "
    f"sum to their sample's outcome difference."
))
```

All four contrasts agree with regression imputation in all three designs; the largest absolute gap is 1e-14. Both decompositions sum to their sample's outcome difference.

8.4 Four population checks of the Kline condition

For an isolated check of the observed-data functional, let $M \mid D = 1$ be uniform on $[-1, 1]$ and set

$$h_1(m) = 1 + 2m + \lambda(m^2 - 1/3),$$
$$r(m) = 1 + 0.6m + \zeta(m^2 - 1/3).$$

Use $\lambda \in \{0, 1.5\}$ and $\zeta \in \{0, 0.9\}$. Each density ratio is positive and integrates to one under the treated distribution. OLS imputation has population value 1.4 in all four cases, while

$$\Theta = 1.4 + \frac{4\lambda\zeta}{45}, \qquad \tilde{\Theta} - \Theta = -\frac{4\lambda\zeta}{45}.$$

The following values use 160-node Gauss–Legendre quadrature. They verify statistical robustness of the functional; no causal interpretation is assumed for this calculation.

Outcome mean	Density ratio	Target	OB	Bias
Linear	Linear	1.400	1.400	+0.000
Linear	Quadratic	1.400	1.400	−0.000
Quadratic	Linear	1.400	1.400	+0.000
Quadratic	Quadratic	1.520	1.400	−0.120

8.5 Sampling uncertainty

For fixed-dimensional, well-behaved regressions, an iid pairs bootstrap can refit the treated regression and recompute both arm averages. It accounts jointly for the estimated response surface and empirical mediator distribution. Resample complete units: resampling mediator components separately would change their joint distribution. Clustered designs require an appropriate cluster bootstrap or influence function calculation.

The illustrative interval below uses 499 pairs resamples of one clean interaction-design sample with 4,000 observations. It is a percentile interval for the pure natural direct effect under the maintained identifying assumptions. It does not protect against hidden mediator–outcome confounding.

```
db, mb, yb = simulate(4000, *designs["Treatment interaction"], rng)
point = ob_fit(db, mb, yb)[0]
draws = np.empty(499)
for j in range(len(draws)):
    ix = rng.integers(0, len(db), size=len(db))
    draws[j] = ob_fit(db[ix], mb[ix], yb[ix])[0]
lower, upper = np.quantile(draws, [0.025, 0.975])
bootstrap = {
    "n": len(db), "resamples": len(draws), "point": float(point),
    "percentile_lower": float(lower), "percentile_upper": float(upper),
    "true_nde0": float(natural_effects(designs["Treatment interaction"][0])[0]),
}
display(Markdown(
    f"The estimate is **{point:.3f}**, with a bootstrap 95% interval "
    f"**[{lower:.3f}, {upper:.3f}]**. "
    f"The structural pure natural direct effect is {bootstrap['true_nde0']:.3f}."
))
```

The estimate is **1.526**, with a bootstrap 95% interval **[1.445, 1.613]**. The structural pure natural direct effect is 1.570.

9 Interpretation

Regression imputation is a direct implementation of the mediation formula. With a treated outcome regression and a control mediator distribution, the imputed level targets Ψ_{10}; the OB unexplained component targets NDE_0 and the explained component targets NIE_1. An interacted pooled regression centered at the control mediator mean gives exactly the same direct-effect estimate. The same fit gives both natural indirect effects by multiplying the mediator mean shift by the relevant arm's outcome slope.

The Baron–Kenny product formula is the common-slope linear special case. Treatment–mediator interactions require a reference mediator distribution, which regression imputation makes explicit. Joint mediator covariance affects statistical precision and conditioning. Mediator–outcome exchangeability and cross-world restrictions determine whether the estimated decomposition has a natural-effect interpretation.

10 Computational supplement

The Quarto source contains every simulation, estimator, quadrature calculation, bootstrap, and check. It runs without saved numerical inputs. The replication archive includes a locked Python environment specification. All numerical results are simulated illustrations rather than empirical estimates from the cited papers.

The deterministic assertions verify the two exact sample decompositions, the centered interacted-regression identity, all four interacted-effect contrasts and both sample decompositions, mediator balancing, pooled whitening invariance, the zero augmentation, and all four population robustness cases. A separate Monte Carlo check compares every estimator's mean to its analytically derived probability limit. The check uses the Monte Carlo standard error of the mean over independent replications; it does not treat biased estimators as accurate estimates of the causal target.

All deterministic checks pass. The largest absolute standardized Monte Carlo deviation from an analytic probability limit is 1.05. The Monte Carlo comparison contains 1,800 estimator fits.

References

Bibliography

[1] K. Imai, L. Keele, and T. Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, vol. 25, no. 1, pp. 51–71, 2010, doi: 10.1214/10-STS321.

[2] E. J. Tchetgen Tchetgen and I. Shpitser, "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness and Sensitivity Analysis," *The Annals of Statistics*, vol. 40, no. 3, pp. 1816–1845, 2012, doi: 10.1214/12-AOS990.

[3] R. M. Baron and D. A. Kenny, "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, vol. 51, no. 6, pp. 1173–1182, 1986, doi: 10.1037/0022-3514.51.6.1173.

[4] W. Lin, "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 295–318, 2013, doi: 10.1214/12-AOAS583.

[5] P. Kline, "Oaxaca–Blinder as a Reweighting Estimator," *American Economic Review*, vol. 101, no. 3, pp. 532–537, 2011, doi: 10.1257/aer.101.3.532.