Mediation analysis has been around a long time, though its popularity has varied between disciplines and over the years. While some fields have been attracted to the potential of mediation models to identify pathways, or mechanisms, through which an independent variable affects an outcome, others have been skeptical that the analysis of mediated relationships can ever be done scientifically.

Two developments, one more scientific than the other, have led to a renewed popularity of mediation analysis. The first, less scientific reason is that software developments have made it easier than ever to fit mediation models. Drag-and-drop interfaces such as AMOS, or the menu-driven PROCESS add-on to SPSS, have meant that researchers can easily test different models – often atheoretically in a \(p\)-hacking exercise – and get results that seem *prima facie* believable. However, it is entirely possible, and very easy, to utilize this software without any understanding of the assumptions mediation analysis requires to yield a causal interpretation. It turns out that many models that have passed peer review and ended up in scientific journals have no causal interpretation whatsoever. Simply drawing boxes and arrows, then hitting “Run” in the software of choice, is not appropriate methodology.

In the second, more scientific development, statisticians from computer science, epidemiology, psychology, and political science have separately, but in a complementary manner, brought mediation into the causal inference framework. This has not only illuminated the assumptions needed to call a mediation model “causal,” but it has also created tools that allow researchers to assess the sensitivity of their causal claims to unobserved confounders.

The newer potential outcomes framework provides concrete definitions for mediated effects, which had been seriously lacking in the prior mediation literature. Does it make sense to study the effect of an intervention while holding the mediator constant? If so, how do we interpret holding the mediator constant when the mediator is affected by other confounders outside the researcher’s control? Or should we hold the treatment constant and determine the effect on outcomes when we vary the mediator, which we might prefer if intervening on the mediator is cheaper than the treatment? The real-world consequences of mediation have until recently been glossed over, but the potential outcomes framework allows us to consider different scenarios.

This blog post will review the history of mediation prior to its incorporation into the potential outcomes framework, when the focus was on determining the appropriate standard error for indirect effects. It then describes how the meaning of an indirect effect is more nuanced than most applications have given it credit for. The subsequent section describes the potential outcomes notation and the assumptions underlying the interpretation of a mediation pathway as causal. The final section provides three examples of when the old-school approach to mediation works and, more importantly, when it fails.

Throughout, the discussion will, unless indicated otherwise, assume a binary intervention, a single continuous mediator, and a continuous outcome that is modeled as a linear function of the intervention, mediator, and (possibly) confounders. This is the simplest and most familiar situation. However, it should become evident that even in this seemingly straightforward scenario there are enough complications to show the reader how much care is required in interpreting mediation models. Moving into the world of categorical outcomes, and especially multiple mediators that affect each other, only makes things more complex.

**A Brief History of Mediation**

Take the following simple and familiar mediation model.

\(X\) is a binary intervention that affects \(Y\) both directly and indirectly by changing the value of the mediator \(M\), which in turn affects \(Y\). This DAG corresponds to the following linear models:

\[  
\begin{align}  
\mathbb{E}[M] &= \gamma\_0 + \gamma\_1X \\  
\mathbb{E}[Y] &= \beta\_0 + \beta\_1M + \beta\_2X  
\end{align}  
\]

Substituting for \(M\) in the model of \(Y\) yields:

\[  
\begin{align}  
\mathbb{E}[Y] &= \beta\_0 + \beta\_1(\gamma\_0 + \gamma\_1X) + \beta\_2X \\  
& = (\beta\_0 + \beta\_1\gamma\_0) + \beta\_1\gamma\_1X + \beta\_2X  
\end{align}  
\]

The substitution makes it clear that the direct effect of \(X\) on \(Y\) can be recovered as the term \(\beta\_2\) from a regression of \(Y\) on \(X\) and \(M\). The indirect effect can be recovered as the product of \(\beta\_1\) from regressing \(Y\) on both \(X\) and \(M\), times \(\gamma\_1\) from the regression of \(M\) on \(X\). Whereas the significance of the direct effect can be found in the usual regression output, the significance of the indirect effect needs an extra step. Three ways of testing the significance have been proposed.

First, an oft cited 1986 article from Baron & Kenny suggested a three-step approach to assessing mediation:

1. Regress \(Y\) on \(X\) only. This first step establishes whether there is any relationship at all that could be mediated.
2. Regress \(M\) on \(X\) only. This establishes that there is a relationship between \(X\) and the mediator. If \(M\) is unresponsive to \(X\), it cannot mediate.
3. Regress \(Y\) on \(X\) and \(M\) simultaneously. If the previously significant relationship between \(Y\) and \(X\) disappears, the relationship is completely mediated. If the previously significant relationship is still significant, but the magnitude is diminished, the association is partially mediated.

This approach still does not provide a standard error or confidence interval for the product term \(\beta\_1\gamma\_1\). Sobel (1982) therefore proposed applying the delta method to the system of equations defined above by the separate regressions for \(Y\) and \(M\). For a \(\theta\) that is a function \(f\) of two random variables \(a\) and \(b\),

\[  
Var\_{\theta} = \nabla f(a,b)^{\prime}cov(a,b) \nabla f(a,b)  
\]  
For the indirect effect we have

\[  
\theta = f(\beta\_1, \gamma\_1) = \beta\_1\gamma\_1  
\]  
The gradient vector is:

\[  
\begin{bmatrix}  
\beta\_1 \\  
\gamma\_1  
\end{bmatrix}  
\]

So our first-order variance approximation is:

\[  
\begin{align}  
Var\_{\gamma\_1\beta\_1} &=  
\begin{bmatrix}  
\beta\_1 \gamma\_1  
\end{bmatrix}  
\begin{bmatrix}  
Var\_{\beta\_1} Cov\_{\beta\_1, \gamma\_1} \\  
Cov\_{\beta\_1, \gamma\_1} Var\_{\gamma\_1}  
\end{bmatrix}  
\begin{bmatrix}  
\beta\_1 \\  
\gamma\_1  
\end{bmatrix} \\  
& = \beta\_1^2Var\_{\beta\_1} + \gamma\_1^2Var\_{\gamma\_1} + 2\beta\_1\gamma\_1Cov\_{\beta\_1, \gamma\_1}  
\end{align}  
\]

When the models for \(M\) and \(Y\) are assumed to be independent, as is typically done, the covariance term is zero. The standard error is consequently estimated by taking the square root of the variance:

\[  
SE\_{\beta\_1\gamma\_1} = \sqrt{\beta\_1^2SE\_{\beta\_1}^2 + \gamma\_1^2SE\_{\gamma\_1}^2}  
\]

where \(SE\_{\beta\_1}\) is the standard error for \(\beta\_1\) taken from the regression of \(Y\) on \(X\) and \(M\), and \(SE\_{\gamma\_1}\) is the standard error of \(\gamma\_1\) from the regression of \(M\) on \(X\). In large samples, the product divided by its Sobel-method standard error is distributed normally and can therefore be used to test for significance.

The Sobel method is implemented in most software packages that perform mediation. However, it is based on an approximation that requires \(N\) to be large enough for the sampling distribution to be normal. Because researchers do not always have the luxury of large samples, the third approach to assessing indirect effects is to rely on the nonparametric bootstrap. This method boils down to treating the sample as though it were a population, repeatedly sampling from the sample with replacement, calculating the product term on each bootstrap sample, and using the standard deviation of the resulting distribution as the standard error.

**The Troubles**

All of this is fine practice assuming the simple mediation model presented in the path diagram above holds. However, once confounders are introduced the situation becomes a bit more murky. And, for observational data, there are always confounders. A simple example taken from Pearl (2014) illustrates two problems created by confounders: direct effect estimates may be biased, and the meaning of direct/indirect effects is ambiguous.

Take the following model:

Here \(X\) does not have any direct or indirect effect on \(Y\) except through the backdoor path \(X \rightarrow M \leftarrow C \rightarrow Y\), where \(C\) affects both \(M\) and \(Y\). Assume \(X\) and \(M\) are both binary and the model equations have the following linear specifications, with coefficients of either 0 or 1:

\[  
\begin{align}  
Y &= 0X + 0M + 1C \\  
M &= 1X + 1C  
\end{align}  
\]  
The second equation written in terms of \(C\) is,

\[  
C = M – X  
\]  
The direct effect estimate of \(X \rightarrow Y\) we would like to recover should equal zero. Yet if we control for \(M\), as per the Baron and Kenny approach, we end up instead with an estimate of \(-1\). Why?

Controlling for \(M\) means that we change \(X\) from zero to one while holding \(M\) constant at, say, zero. In this scenario, with \(X = 0\) and \(M = 0\), we get:

\[  
\begin{align}  
Y &= C \\  
&= M – X \\  
&= 0 – 0 \\  
&= 0  
\end{align}  
\]

When we switch to \(X = 1\), still holding \(M\) constant at zero, we get:

\[  
\begin{align}  
Y &= C \\  
&= M – X \\  
&= 0 – 1 \\  
&= -1  
\end{align}  
\]

Thus, changing \(X\) from 0 to 1 while holding \(M\) constant gets an effect of \(0 – 1 = -1\).

\(M\) is a function of both \(X\) and \(C\), so holding \(M\) constant at zero would require \(C\) to also change in order to maintain the equality \(M = X+C\). To keep the structural model’s integrity, we would need to introduce a *counterfactual* world in which \(M\) is forced to take on the value of zero regardless of \(C\). This is equivalent to removing the \(C \rightarrow M\) path, which is not the same as statistically controlling for \(M\).

Defining direct and indirect effects therefore needs to be based carefully on what we think we can intervene on given the model’s structural relationships. Are we interested in the effect of changing \(X\) from, say, \(X = 1\) to \(X = 0?\) If so, we need to understand what “holding \(M\) constant” means given the presence of confounders. Or are we interested in the effect of changing \(M\), say from \(M = 1\) to \(M = 0\)? Are we interested in both? Switching to a counterfactual framework will force us to be more explicit about what we are manipulating and how we report direct and indirect effects. It also forces us to be explicit in the assumptions necessary to *identify* (estimate from data) these effects.

The next section turns to defining the different types of direct effects we can estimate along with more concrete definitions.

**Potential Outcomes and Mediation**

The counterfactual approach to treatment effects is now well-established for non-mediation models. The idea is that every subject has multiple *potential outcomes*, one that occurs if the treatment is received \((X = 1)\) and one that occurs if the treatment is not received \((X = 0)\). We refer to a single subject’s two potential outcomes as \(Y\_0\) for the outcome when \(X = 0\) and \(Y\_1\) for the outcome when \(X = 1\). If we observed both of these, the treatment effect for this subject could be calculated as \(Y\_1 – Y\_0\). However, we only observe a treated subject receiving the treatment; the non-treatment is the unobserved counterfactual. Alternatively, we only observe the non-treated subject not receiving the treatment; the treatment outcome is the unobserved counterfactual. While we cannot identify the individual-specific treatment effect, it is easy to show that we *can* estimate the *Average Treatment Effect* (ATE) across all individuals in our sample.

Mediation requires an expansion of the potential outcomes. We first have two counterfactuals for the mediator \(M\) given treatment \(X\):

1. The value of \(M\) when \(X = 0\), \(M\_0\).
2. The value of \(M\) when \(X = 1\), \(M\_1\).

We have four counterfactuals for the outcome \(Y\).

1. The value of \(Y\) when \(X = 0\) and \(M = M\_0\), \(Y\_{0M\_0}\).
2. The value of \(Y\) when \(X = 1\) and \(M = M\_1\), \(Y\_{1M\_1}\).
3. The value of \(Y\) when \(X = 0\) and \(M = M\_1\), \(Y\_{0M\_1}\).
4. The value of \(Y\) when \(X = 1\) and \(M = M\_0\), \(Y\_{1M\_0}\).

Of these four, only one is observed for a given individual, and the third and fourth will never be observed for any individual. The latter two may seem odd to consider at first glance. However, given these potential outcomes, we can state the following definitions (Pearl, 2001):

1. The *controlled direct effect* (CDE): The change in \(Y\) when switching from \(X = 0\) to \(X = 1\) if \(M\) were set at the same value for everybody.
2. The *natural direct effect* (NDE): The change in \(Y\) when switching from \(X = 0\) to \(X = 1\) if \(M\) were set at \(M\_0\) for everybody. That is, the NDE is the direct effect if the mediator were forced to take on the value it would have, for each individual, in the absence of treatment (\(X = 0\)).
3. The *natural indirect effect* (NIE): The change in \(Y\) when keeping \(X = 1\) for everybody and changing \(M = 0\) to \(M = 1\).

Using counterfactual notation:

1. CDE: \(Y\_{1M\_{m}} – Y\_{0M\_{m}}\)
2. NDE: \(Y\_{1M\_0} – Y\_{0M\_0}\)
3. NIE: \(Y\_{0M\_1} – Y\_{0M\_0}\)

Note that the \(M\_m\) subscript for the CDE means that the value of the CDE will change depending on the value we set for \(M\).

In classical mediation modeling, the “total” effect of an intervention is partitioned into the direct effect plus the indirect effect. The NIE and NDE are defined so that the total effect \(TE = NIE + NDE\). In the special case of a linear model with no treatment by mediator interaction and no post-treatment confounders, a CDE can equal the NDE. When this special case does not obtain, the CDE will be different from the NDE and cannot be added to the NIE to get the total effects. Nonetheless, the benefit of the CDE, as the next section discusses, is that it requires fewer assumptions to estimate. Thus, the direct causal effect of an intervention under specific values of the mediator can be recovered even in cases when the causal *indirect* effect is not identified.

Being based on unobserved counterfactuals, all three types of effects are not identified at the individual level. As with ATEs in traditional causal inference, we instead focus on the *average* CDE, NDE, and NIE, as these effects can be identified from the data we do observe given certain assumptions are met. The next section turns to describing these assumptions and how they lead to identification of direct and indirect effects.

**Assumptions**

As a point of reference, consider the following mediation model with confounders:

Here \(X\) is the treatment, \(Y\) is the outcome, and \(M\) is the mediator. \(C\_1\), \(C\_2\), and \(C\_3\) are vectors of confounders, which may or may not contain the same variables. From Vanderweele (2015), we will make the following assumptions:

* A1: There is no unmeasured confounding of the treatment-outcome relationship.
* A2: There is no unmeasured confounding of the mediator-outcome relationship.
* A3: There is no unmeasured confounding of the treatment-mediator relationship.
* A4: There is no mediator-outcome confounder that is affected by the exposure.

Based on the figure, these assumptions correspond to the following:

* A1: \(C\_1\) contains all confounders of the \(X \rightarrow Y\) path, is observed, and is included in the model of \(Y\).
* A2: \(C\_2\) contains all confounders of the \(M \rightarrow Y\) path, is observed, and is included in the model of \(Y\).
* A3: \(C\_3\) contains all confounders of the \(X \rightarrow M\) path, is observed, and is included in the model of \(M\).
* A4: The only directed path from \(X\) to \(M\) does not pass through any intermediate variables. This would be violated, for example, if the path \(X \rightarrow C\_2 \rightarrow M\) were to exist.

The (average) natural indirect effect is defined as:

\[  
NIE = \mathbb{E}[Y\_{0M\_1} – Y\_{0M\_0}\mid C] = \mathbb{E}[Y\_{0M\_1} \mid C] – \mathbb{E}[Y\_{0M\_0} \mid C]  
\]

We thus have two terms, \(\mathbb{E}[Y\_{0M\_1} \mid c]\) and \(\mathbb{E}[Y\_{0M\_0} \mid c]\) that we hope to be able to write in terms of our observed data. Pearl (2001) provides proofs, based on standard expectation algebra, of how these assumptions lead to the identification of each term from observable data.

\[  
\begin{align}  
(A)NIE &= \mathbb{E}[Y\_{0M\_1} \mid c] – \mathbb{E}[Y\_{0M\_0} \mid c] \\  
&= \sum\_m \mathbb{E}[Y \mid x = 0, m, c] \{P(m \mid x = 0, c) – P(m \mid x = 1, c)\}  
\end{align}  
\]

This estimand is identified because all of the terms on the RHS of the equation can be recovered from the data.

Pearl’s notation may be unfamiliar to researchers who have traditionally worked in the regression-based Baron and Kenny world. Vanderweele and Vansteelandt (2009) show how regression can be used to recover the corresponding terms. Assuming that all four assumptions are met and the correct functional form is linear, we can specify two equations:

\[  
\begin{align}  
\mathbb{E}(M \mid X = x, M = m, C = c) &= \theta\_0 + \theta\_1x + \theta\_2m + \theta\_3xm + \theta\_4^{\prime}c \\  
\mathbb{E}(Y \mid X = x, M = m, C = c) &= \beta\_0 + \beta\_1x + \beta\_2^{\prime}c  
\end{align}  
\]

where \(c\) is a vector of confounders and \(xm\) is a possible interaction between the treatment and mediator. Given a change in treatment from \(x\) to \(x^\*\), the treatment effects from a linear structural model are:

\[  
\begin{align}  
CDE(m) &= (\theta\_1 + \theta\_3m(x – x^\*)) \\  
NDE &= (\theta\_1 + \theta\_3\beta\_0 + \theta\_3\beta\_1x^\* + \theta\_3\beta\_2^{\prime}c)(x – x^\*) \\  
NIE &= (\theta\_2\beta\_1 + \theta\_3\beta\_1x)(x – x^\*)  
\end{align}  
\]

As the examples below will show, these formulas reduce to the Baron and Kenny approach if there is no treatment by mediator interaction \((\theta\_3 = 0)\) and all assumptions are met. Note, though, that Vanderweeele (2015) shows that it is generally not a good idea to assume no interaction is present. This is because:

1. The interaction term allows for a fuller picture of the causal mechanism.
2. Statistical tests that show non-significance of the interaction should not be trusted due to low power, which is often the case for interactions.
3. Even if the interaction term is not significant in a regression, its inclusion can sometimes boost the power to identify other significant effects.

The assumptions A1-A4 deal with the ability to measure and control for confounders. While this is no different from any regression model, the problem is made more complicated by the fact that confounders can affect multiple paths. Omitted variable bias is consequently a bigger threat to the validity of mediation estimates than traditional single-equation regression models.

Interestingly, the multi-equation set-up of mediation modeling provides a unique opportunity to both test the robustness of results to unmeasured confounders and quantify just how large the effect of those confounders would have to be to explain away the causal effect. The calculations are based on the observation that, in the absence of confounding, the errors from the mediation model and the outcome model should be uncorrelated (Imai, Keele, and Yamamoto, 2010). Tingley et al. (2014) have developed the R package mediation to perform the sensitivity analysis and present useful graphical summaries of the results. Vanderweele (2015) discusses an alternative sensitivity test that can also be applied to single equation models. Describing these in greater detail is beyond the scope of this already lengthy post, but the reader should consult the publications, as sensitivity analyses are becoming standard practice in mediation.

Sensitivity analyses are great for testing the no-confounding assumptions A1-A3, but assumption 4 (no confounder of \(M \rightarrow Y\) is affected by \(X\)) may be very strong. Indeed, it rules out using regression for models like the following:

It is well known that conditioning an a post-treatment covariate can induce spurious correlations and should be avoided (Acharya, Blackwell, and Sen, 2016). Yet not controlling for \(C\) means that the estimate for \(M\_1 \rightarrow Y\) will be biased. The upshot is that traditional regression modeling cannot be used to partition total effects into separate indirect and direct effects that have a causal meaning.

A variation on this is the case when there are multiple, non-independent mediators, as in the following:

We label \(M\_2\) now as a second mediator rather than a confounder because it is of theoretical interest. If there were no \(M\_2 \rightarrow M\_1\) path, we could still proceed using regression modeling and calculating separate NIEs. However, the presence of the \(M\_2 \rightarrow M\_1\) complicates things, and standard regression is either not an option or requires additional assumptions. This is discussed in several publications (see, for example, Imai and Yamamoto, 2013, and citations therein).

On the other hand, it turns out that the CDE only requires assumptions 1 and 2. In the notation of Pearl’s SCMs, we avoid the issues of a post-treatment confounder of \(M \rightarrow Y\) by simply disabling any arrows pointing at \(M\) and setting its value to a constant, as though we were intervening on the observational measure ourselves. Though we only intervene in theory, the result of doing so is an estimate of the direct effect conditional on the value of \(M\).

\[  
CDE = P(Z=z \mid do(X = x), do(M = m)) – P(Z=z \mid do(X = x^{\prime}), do(M = m))  
\]

This result may seem unsatisfying for somebody interested in a statistical test of the indirect effect. However, as Acharya et al. (2016) discuss, the CDE can sometimes illuminate causal mechanisms on its own. First, a nonzero estimate of the CDE rules out complete mediation by the mediator. That is, there is some causal path other than the one that passes through the proposed mediator. Second, if we assume, as is often done, that there is no interaction between the treatment and the mediator, the CDE and the NDE are equal. This means that one can recover the NIE by subtracting out the CDE from the total effect.

Recall, however, that Vanderweele (2015) argues in favor of always including such an interaction effect. In addition, although identified, regression cannot always be used to estimate the CDE. If assumption 4 is in fact violated, the effect is still identified, but a different approach to estimation is required (Vansteelandt, 2009), or much stronger assumptions are necessary (Imai & Yamamoto, 2013).

**Example 1**

We’ll demonstrate the problems under three different scenarios. First, we’ll look at a case where the assumptions can be satisfied with appropriate controls and where there’s no interaction between the treatment and mediator.

Here there are two confounders. \(C\_1\) confounds the association between \(X\) and \(Y\), and \(C\_2\) confounds the association between \(M\) and \(Y\). Without proper control, these violate assumptions 1 and 2. First, we’ll generate some data consistent with the model:

\[  
\begin{align}  
\mathbb{E}(M \mid X = x, M = m, C\_1 = c\_1) &= \theta\_0 + \theta\_1x + \theta\_2m + \theta\_3c\_1 + \theta\_4c\_2 \\  
\mathbb{E}(Y \mid X = x, M = m, C\_1 = c\_1, C\_2 = c\_2) &= \beta\_0 + \beta\_1x + \beta\_2c\_1  
\end{align}  
\]

set.seed(12345)

C1 <- rnorm(10000) # Generate first random confounder

X <- rbinom(10000, 1, plogis(.8\*C1)) # Generate random tx variable as a function of C1

C2 <- rnorm(10000) # Generate second random confounder

M <- rnorm(10000, .8\*X + .8\*C2) # Generate mediator as function of tx and second confounder

Y <- rnorm(10000, .8\*X + .8\*M + .8\*C1 + .8\*C2) # Model outcome

tbl <- tibble(

C1 = C1,

C2 = C2,

X = X,

M = M,

Y = Y)

The data generating process assumed an NDE of 0.80 and an NIE of 0.64. Let’s fit these models:

mod\_y <- lm(Y ~ X + M + C1 + C2, data = tbl) %>%

broom::tidy() %>%

mutate\_if(is.numeric, ~round(.x, 3)) %>%

mutate(p.value = if\_else(

p.value < 0.001,

true = "< 0.001",

false = as.character(round(p.value, 3))))

mod\_m <- lm(M ~ X + C2, data = tbl) %>%

broom::tidy() %>%

mutate\_if(is.numeric, ~round(.x, 3)) %>%

mutate(p.value = if\_else(

p.value < 0.001,

true = "< 0.001",

false = as.character(round(p.value, 3))))

kable(mod\_y, align = c("l", rep("c", 4)))

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.015 | 0.015 | 1.043 | 0.297 |
| X | 0.794 | 0.023 | 34.549 | < 0.001 |
| M | 0.810 | 0.010 | 78.972 | < 0.001 |
| C1 | 0.808 | 0.011 | 74.832 | < 0.001 |
| C2 | 0.795 | 0.013 | 60.938 | < 0.001 |

kable(mod\_m, align = c("l", rep("c", 4)))

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.017 | 0.014 | 1.199 | 0.231 |
| X | 0.767 | 0.020 | 38.988 | < 0.001 |
| C2 | 0.806 | 0.010 | 81.819 | < 0.001 |

Since there is no interaction, \(\theta\_3 = 0\), the NDE formula is:

\[  
\begin{align}  
NDE &= (\theta\_1)(x – x^\*) \\  
&= (0.794)(1 – 0) \\  
&= 0.794  
\end{align}  
\]

The estimate of the NIE is

\[  
\begin{align}  
NIE &= (\theta\_2\beta\_1)(x – x^\*)) \\  
&= (0.81 \times 0.767)(1 – 0)) \\  
&= 0.62127  
\end{align}  
\]

What do we get from the Baron and Kenny method? We can use the lavaan package to return the direct and indirect effects the usual way. The following are three incompletely specified models and one correct specification.

library(lavaan)

## Warning: package 'lavaan' was built under R version 3.6.1

# Moodel without confounders

mod\_1 <- "Y ~ a\*X + b\*M

M ~ X

ind\_fx := a\*b"

# Model with confounder 1

mod\_2 <- "Y ~ a\*X + b\*M + C1

M ~ X

ind\_fx := a\*b"

# Model with confounder 2

mod\_3 <- "Y ~ a\*X + b\*M + C2

M ~ X + C2

ind\_fx := a\*b"

# Model with both confounders

mod\_4 <- "Y ~ a\*X + b\*M + C1 + C2

M ~ X + C2

ind\_fx := a\*b"

Here’s a function to pull out just the direct and indirect effects estimates.

get\_fx <- function(df, model) {

mod <- sem(model, data = df) %>%

parameterEstimates() %>%

filter(label %in% c("a", "ind\_fx")) %>%

as\_tibble

tibble(Effect = c("NDE", "NIE"),

Estimate = pull(mod, est),

p.value = pull(mod, pvalue)) %>%

mutate(p.value = if\_else(

p.value < 0.001,

true = "< 0.001",

false = as.character(round(p.value, 3)))) %>%

mutate(Estimate = round(Estimate, 3))

}

Map over the models:

list(`No Confounders` = mod\_1,

`Confounder 1` = mod\_2,

`Confounder 2` = mod\_3,

`Both Confounders` = mod\_4) %>%

map\_dfr(~get\_fx(tbl, .x), .id = "Model Type") %>%

arrange(Effect) %>%

kable()

| **Model Type** | **Effect** | **Estimate** | **p.value** |
| --- | --- | --- | --- |
| No Confounders | NDE | 1.086 | < 0.001 |
| Confounder 1 | NDE | 0.508 | < 0.001 |
| Confounder 2 | NDE | 1.373 | < 0.001 |
| Both Confounders | NDE | 0.794 | < 0.001 |
| No Confounders | NIE | 1.310 | < 0.001 |
| Confounder 1 | NIE | 0.612 | < 0.001 |
| Confounder 2 | NIE | 1.115 | < 0.001 |
| Both Confounders | NIE | 0.643 | < 0.001 |

Clearly, appropriate control needs to be made for confounders. But you knew that, so let’s make this a little more complicated.

**Example 2**

We’ll take the same model as before, but now we’ll violate assumption 4 by introducing an association between the treatment and the mediator-outcome confounder:

The only change is that there is now one additional path, \(X \rightarrow C\_2\). We’ll generate the data:

C1 <- rnorm(10000) # Generate first random confounder

X <- rbinom(10000, 1, plogis(.8\*C1)) # Generate random tx variable as a function of C1

C2 <- rnorm(10000, .8\*X) # Generate second confounder as function of tx

M <- rnorm(10000, .8\*X + .8\*C2) # Generate mediator as function of tx and second confounder

Y <- rnorm(10000, .8\*X + .8\*M + .8\*C1 + .8\*C2) # Model outcome

tbl <- tibble(

C1 = C1,

C2 = C2,

X = X,

M = M,

Y = Y)

The data generating process is again consistent with an NIE 0.64. The NDE is now 1.44, as it includes both the direct path of \(X \rightarrow Y\) (.8) and the path \(X \rightarrow C\_2 \rightarrow Y\) (.8\*.8 = .64).

Fit the models the traditional way using lavaan.

# Moodel without confounders

mod\_1 <- "Y ~ a\*X + b\*M

M ~ X

ind\_fx := a\*b"

# Model with confounder 1

mod\_2 <- "Y ~ a\*X + b\*M + C1

M ~ X

ind\_fx := a\*b"

# Model with confounder 2

mod\_3 <- "Y ~ a\*X + b\*M + C2

M ~ X + C2

ind\_fx := a\*b"

# Model with both confounders

mod\_4 <- "Y ~ a\*X + b\*M + C1 + C2

M ~ X + C2

ind\_fx := a\*b"

Map over models.

list(`No Confounders` = mod\_1,

`Confounder 1` = mod\_2,

`Confounder 2` = mod\_3,

`Both Confounders` = mod\_4) %>%

map\_dfr(~get\_fx(tbl, .x), .id = "Model Type") %>%

arrange(Effect) %>%

kable()

| **Model Type** | **Effect** | **Estimate** | **p.value** |
| --- | --- | --- | --- |
| No Confounders | NDE | 1.479 | < 0.001 |
| Confounder 1 | NDE | 0.919 | < 0.001 |
| Confounder 2 | NDE | 1.392 | < 0.001 |
| Both Confounders | NDE | 0.823 | < 0.001 |
| No Confounders | NIE | 1.758 | < 0.001 |
| Confounder 1 | NIE | 1.096 | < 0.001 |
| Confounder 2 | NIE | 1.127 | < 0.001 |
| Both Confounders | NIE | 0.664 | < 0.001 |

Recall that the NDE is the change in \(Y\) when we change from \(X = x\) to \(X = x^\*\), holding \(M\) at \(M\_0\) for everybody. When we set \(M = M\_0\) for everybody, what is left is the effect of \(X\) not operating through the mediator. As seen in the table, we get a direct effect estimate of \(B\_{x \rightarrow y} = .823\) when we control for both confounders. Although this is consistent with how we generated the \(M \rightarrow Y\) path, it is *not* the NDE. By the definition of the NDE (which we defined so that NDE + NIE = TE), the estimate should also include the path \(X \rightarrow C\_2 \rightarrow Y\). In other words, the estimate is not consistent with our definition and cannot be added to the NIE to get the total effect.

So, you think, we can just estimate a different model without \(C\_2\) to get the NDE. But, as the table shows, now the estimate of the NIE is wrong. This is a classic case of omitted variable bias because the effect through the confounder is being added to the NIE estimate. (For what it’s worth, the NDE is also wrong.)

There is one final approach that may save us and still allow us to get both the NIE and NDE. The NIE was correct in the model that controlled for both \(C\_1\) and \(C\_2\), we just didn’t get the correct NDE. We could, in theory, specify the SEM to also estimate the regression of \(C\_2\) on \(X\) and calculate the indirect effect of \(X\) on \(Y\) through \(C\_2\). We then just add the \(X \rightarrow Y\) and \(X \rightarrow C\_2 \rightarrow Y\) paths together to get the NDE.

# Model with both confounders

mod\_5 <- "Y ~ a\*X + b\*M + C1 + c\*C2

M ~ X + C2

C2 ~ d\*X

ind\_fx\_1 := a\*b

ind\_fx\_2 := c\*d

nde := a + c\*d"

parameterEstimates(sem(mod\_5, tbl)) %>%

filter(label == "nde") %>%

dplyr::select(label, estimate = est, se, pvalue) %>%

mutate\_if(is.numeric, ~round(.x, 3)) %>%

mutate(pvalue = if\_else(pvalue <= 0, "< .001", as.character(pvalue))) %>%

kable

| **label** | **estimate** | **se** | **pvalue** |
| --- | --- | --- | --- |
| nde | 1.467 | 0.031 | < .001 |

*Now* we have the correct answer. But notice a few things:

1. The NDE did not come out of the regression results, as is usually assumed, but had to be calculated separately on the basis of the SEM that included a \(X \rightarrow C\_2\) path.
2. With observational data, \(C\_2\) is almost certainly *not* a single variable but rather a vector of many variables. The manual process quickly grows tedious.
3. If \(C\_2\) contains variables that cannot be observed, the NIE will be biased.
4. If \(C\_2\) is actually a second mediator, the potential outcomes become much more laborious to derive. The upshot will be that the assumptions also become more onerous, and the model may not be identified for other parametric forms.

**Example 3**

For the final example, we’ll return to the model from example 1. That is, we’ll remove the \(X \rightarrow C\_2\) path so we are not in the unfortunate situation of having a post-treatment confounder to contend with. However, we will introduce an interaction between the treatment and the mediator. The following code chunk gets us this model:

C1 <- rnorm(10000) # Generate first random confounder

X <- rbinom(10000, 1, plogis(.8\*C1)) # Generate random tx variable as a function of C1

C2 <- rnorm(10000, .8\*X) # Generate second random confounder

M <- rnorm(10000, .8\*X + .8\*C2) # Generate mediator as function of tx and second confounder

Y <- rnorm(10000, .8\*X + .8\*M + .8\*X\*M + .8\*C1 + .8\*C2) # Model outcome with x by m interaction

tbl <- tibble(

C1 = C1,

C2 = C2,

X = X,

M = M,

Y = Y) %>%

mutate(MX = M\*X) # For the lavaan software, need to create interaction variable explicitly

The NIE and NDE are now ambiguous because the effect of \(M\) on \(Y\) depends on \(X\), and the effect of \(X\) on \(Y\) depends on \(M\).

We can fit our models via regression and use the Vanderweele (2015) formulas to recover the estimates consistent with our definitions.

mod\_y <- lm(Y ~ X\*M + C1 + C2, data = tbl) %>%

broom::tidy() %>%

mutate\_if(is.numeric, ~round(.x, 3)) %>%

mutate(p.value = if\_else(

p.value < 0.001,

true = "< 0.001",

false = as.character(round(p.value, 3))))

mod\_m <- lm(M ~ X + C2, data = tbl) %>%

broom::tidy() %>%

mutate\_if(is.numeric, ~round(.x, 3)) %>%

mutate(p.value = if\_else(

p.value < 0.001,

true = "< 0.001",

false = as.character(round(p.value, 3))))

kable(mod\_y, align = c("l", rep("c", 4)))

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | -0.001 | 0.015 | -0.073 | 0.941 |
| X | 0.818 | 0.027 | 30.308 | < 0.001 |
| M | 0.800 | 0.013 | 60.976 | < 0.001 |
| C1 | 0.798 | 0.011 | 74.831 | < 0.001 |
| C2 | 0.794 | 0.013 | 61.205 | < 0.001 |
| X:M | 0.789 | 0.016 | 49.258 | < 0.001 |

kable(mod\_m, align = c("l", rep("c", 4)))

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.005 | 0.014 | 0.374 | 0.708 |
| X | 0.811 | 0.021 | 38.042 | < 0.001 |
| C2 | 0.789 | 0.010 | 79.132 | < 0.001 |

If we use the correct formulas, we get the following estimate for the NDE:

\[  
\begin{align}  
NDE &= (\theta\_1 + \theta\_3\beta\_0 + \theta\_3\beta\_1x^\* + \theta\_3\beta\_2c\_1)(x – x^\*) \\  
&= (0.818 + 0.789 \times 0.005 + 0.789 \times 0.811 \times 0 + 0.789 \times 0.789 \times 0)(1 – 0) \\  
&= 0.821945  
\end{align}  
\]

The estimate of the NIE is

\[  
\begin{align}  
NIE &= (\theta\_2\beta\_1 + \theta\_3\beta\_1x(x – x^\*)) \\  
&= (0.8 \times 0.811 + 0.789 \times 0.811 \times 1)(1 – 0)) \\  
&= 1.288679  
\end{align}  
\]

But if we use the Baron and Kenny method, we get the following:

# Model without confounders

mod\_1 <- "Y ~ a\*X + b\*M

M ~ X

ind\_fx := a\*b"

# Model with confounder 1

mod\_2 <- "Y ~ a\*X + b\*M + C1

M ~ X

ind\_fx := a\*b"

# Model with confounder 2

mod\_3 <- "Y ~ a\*X + b\*M + C2

M ~ X + C2

ind\_fx := a\*b"

# Model with both confounders

mod\_4 <- "Y ~ a\*X + b\*M + C1 + C2

M ~ X + C2

ind\_fx := a\*b"

# Model with both confounders & interaction

mod\_5 <- "Y ~ a\*X + b\*M + MX + C1 + C2

M ~ X + C2

ind\_fx := a\*b"

list(`No Confounders` = mod\_1,

`Confounder 1` = mod\_2,

`Confounder 2` = mod\_3,

`Both Confounders` = mod\_4,

`Both Confounders + Interaction` = mod\_5) %>%

map\_dfr(~get\_fx(tbl, .x), .id = "Model Type") %>%

arrange(Effect) %>%

kable()

| **Model Type** | **Effect** | **Estimate** | **p.value** |
| --- | --- | --- | --- |
| No Confounders | NDE | 2.033 | < 0.001 |
| Confounder 1 | NDE | 1.444 | < 0.001 |
| Confounder 2 | NDE | 1.950 | < 0.001 |
| Both Confounders | NDE | 1.371 | < 0.001 |
| Both Confounders + Interaction | NDE | 0.818 | < 0.001 |
| No Confounders | NIE | 3.223 | < 0.001 |
| Confounder 1 | NIE | 2.304 | < 0.001 |
| Confounder 2 | NIE | 2.311 | < 0.001 |
| Both Confounders | NIE | 1.653 | < 0.001 |
| Both Confounders + Interaction | NIE | 0.654 | < 0.001 |

Things worked out so that our NDE estimate is close to the regression result in the model that includes the interaction. However, the NIE is way off when compared to the correct formula presented above. To accurately account for mediators that interact with the treatment, we cannot simply multiply two coefficients together.

**Conclusion**

This post summarizes the history of mediation analysis, including modern approaches based on the potential outcomes framework. Although the examples were based on simple, linear models with a binary treatment, it should be apparent that traditional applications of mediation modeling can fail without considering the assumptions underlying the model. Researchers commonly draw all kinds of boxes and arrows, hit Run in their software, and report the results. The Baron and Kenny and related approaches require very simple models that fully account for all confounders, do not contain interactions, and do not have multiple mediators affecting each other. When the careless modeler starts drawing arrows from any exogenous variable to multiple intermediate variables, the results may easily turn out to be meaningless.

**Citations**

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3), 512-529.

Baron, RM and Kenny, DA. (1986). The Moderator-Mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 51-71.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2), 141-171.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, 411-420. Morgan Kaufmann Publishers Inc..

Pearl J. (2014) Reply to commentary by Imai, Keele, Tingley, and Yamamoto, concerning causal mediation analysis. *Psycholgocal Methods* 19: 488–492

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13, 290-312.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5),  
1-38.

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction.* New York: Oxford University Press.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4), 457-468.

Vansteelandt, S. (2009). Estimating direct effects in cohort and case–control studies. *Epidemiology*, 851-860.