A short note on mediation

Yen-Chi Chen
University of Washington
April 1, 2020

This note is based on the following paper:


(Effect) Mediation is an interesting problem in the causal inference. Consider three random variables $T, M, Y$ and we are interested in the causal effect from treatment $T$ on outcome $Y$. To simplify the problem, consider a binary treatment scenario $T \in \{0, 1\}$. Assume that there is no other confounders. The variable $M$ is called a mediator if there are arrows $T \to M$ and $M \to Y$.

Suppose that there is also an arrow from $T \to Y$. This arrow will be referred to as the direct effect from $T$ on $Y$ and the (directed) path $T \to M \to Y$ represents the indirect effect from $T$ on $Y$.

To illustrate the difference between direct effect and indirect effects, consider the structural equation model problem:

\[
T = \epsilon_T
\]
\[
M = \alpha T + \epsilon_M
\]
\[
Y = \beta T + \gamma M + \epsilon_Y
\]

where $(\epsilon_T, \epsilon_M, \epsilon_Y)$ are independent noises. In this case, we can rewrite $Y$ as

\[
Y = \beta T + \gamma (\alpha T + \epsilon_M) + \epsilon_Y = (\beta + \alpha \gamma) T + \epsilon_Y.
\]

Thus, the regression coefficient between $T$ and $Y$ will be $\beta + \alpha \gamma$, which is sometime called the total effect, which represents the effect from changing $T$ (while allowing $M$ to change with respect to $T$) on $Y$. The slope $\beta$ is the direct effect, it represents the effect from changing $T$ on $Y$ but keep $M$ fixed. The indirect effect $\alpha \gamma$ will be the difference between the total effect and the direct effect. Note that using the do operator, the total effect of $T$ on $Y$ will be

\[
\mathbb{E}(Y|\text{do}(T = 1)) - \mathbb{E}(Y|\text{do}(T = 0))
\]

and you can easily verify that the above expectation gives a value of $(\beta + \alpha \gamma)$.

When we are not using the structural equation models, the direct effect and indirect effect is not so easy to distinguish. Here we use the counterfactual model approach to define them For each $T = t$, we can define a counterfactual variable $M_t$ of the mediator. For the outcome variable $Y$, its counterfactual is $Y_{t,m}$ since any pair of $T = t$ and $M = m$ could create a potential outcome. Variable $M_t = M|\text{do}(T = t)$ and variable $Y_{t,m} = Y|\text{do}(T = t, M = m)$. Thus, if we only control $\text{do}(T = t)$, then the potential outcome will be $Y_{t,M_t} = Y|\text{do}(T = t)$. 

With these variables, the total effect is
\[
\text{TE} = \mathbb{E}(Y|\text{do}(T = 1)) - \mathbb{E}(Y|\text{do}(T = 0)) \\
= \mathbb{E}(Y_{1,M_t}) - \mathbb{E}(Y_{0,M_0}).
\]  

(1)

The direct effect would be dependent on \( M = m \) since different mediator’s value could lead to a different direct effect from \( T \) on \( M \). It turns out that there could be different versions of direct effects. The first one is the \textit{controlled direct effect}:
\[
\text{CDE}(m) = \mathbb{E}(Y|\text{do}(T = 1, M = m)) - \mathbb{E}(Y|\text{do}(T = 0, M = m)) \\
= \mathbb{E}(Y_{1,m}) - \mathbb{E}(Y_{0,m}).
\]  

(2)

It is the direct effect if we control \( M = m \).

Instead of controlling \( M \), we may allow it to vary according to some distribution. When we let \( M \) to be the random variable following the untreated case \( T \), i.e., \( T = 0 \), this will the \textit{natural direct effect}:
\[
\text{NDE} = \mathbb{E}(Y|\text{do}(T = 1); M \sim M_0) - \mathbb{E}(Y|\text{do}(T = 0); M \sim M_0) \\
= \mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_{0,M_0}).
\]  

(3)

For the indirect effect, there are two versions of it. The first one is the \textit{natural indirect effect}:
\[
\text{NIE} = \mathbb{E}(Y|\text{do}(T = 0); M \sim M_1) - \mathbb{E}(Y|\text{do}(T = 0); M \sim M_0) \\
= \mathbb{E}(Y_{0,M_1}) - \mathbb{E}(Y_{0,M_0}).
\]  

(4)

Namely, \( \text{NIE} \) measures the indirect effect from \( T \) on \( M \) and then on \( Y \) while holding the direct effect as \( T = 0 \). Similarly, we can defined the \textit{treated indirect effect}:
\[
\text{TIE} = \mathbb{E}(Y|\text{do}(T = 1); M \sim M_1) - \mathbb{E}(Y|\text{do}(T = 1); M \sim M_0) \\
= \mathbb{E}(Y_{1,M_1}) - \mathbb{E}(Y_{1,M_0}),
\]  

(5)

which measures the indirect effect on the treated \( (T = 1) \) case.

Here is an interesting property. Let
\[
\text{NDE} = \mathbb{E}(Y_{0,M_1}) - \mathbb{E}(Y_{1,M_1}), \quad \text{NIE} = \mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_{1,M_1}), \quad \text{TIE} = \mathbb{E}(Y_{0,M_0}) - \mathbb{E}(Y_{0,M_1}),
\]
as the case where we swap 0 and 1. Then the total effect can be written as
\[
\text{TE} = \mathbb{E}(Y_{1,M_1}) - \mathbb{E}(Y_{0,M_0}) = \text{NDE} - \text{NIE} = \text{TIE} - \text{NDE}.
\]

1 Identification

To identify the above 5 different types of effects, we need to make some assumptions. The total effect is easiest to be identified. We only need to assume
(A1) \( p(Y = y|\text{do}(T = t)) \) is identifiable.

This assumption will be true if we are in a randomized trial or in an observation study where we have controlled all confounding variables.

However, identification of the indirect effect is more challenging. Consider the following three assumptions:

(A2) \( M_t, Y_{t'}, m \) are independent for all \( t, t', m \).

(A3) \( p(M = m|\text{do}(T = t)) \) is identifiable.

(A4) \( p(Y = y|\text{do}(T = t, M = m)) \) is identifiable.

Note that assumptions (A2-4) also implies the identification of the total effect.

The key to the identification is that we need to be able to compute the distribution of \( Y_t, M_t' \) for all \( t, t' \). To see how (A2-4) identifies the distribution of \( Y_t, M_t' \), we can write its density as

\[
p(Y_{t,M_t'} = y) = \int p(Y_{t,M_t'} = y, M_{t'} = m) dm
\]

\[
= \int p(Y_{t,M_t'} = y|M_{t'} = m)p(M_{t'} = m) dm
\]

\[
= \int p(Y_{t,m} = y|M_{t'} = m)p(M_{t'} = m) dm
\]

\[
= \int p(Y_{t,m} = y)p(M_{t'} = m) dm
\]

\[
= \int p(Y = y|\text{do}(T = t, M = m))p(M = m|\text{do}(T = t'))dm.
\]

(A3) and (A4) identifies the two distributions, so we can identify the entire distribution of \( Y_{t,M_t'} \). In fact, we may relax the assumption (A4) by assuming that \( p(Y_{t,m} = y|M_{t'} = m) \) is identifiable from the above equality.

Using the above form, we can interpret the distribution of \( Y_{t,M_t'} \) as the weighted version of the distribution of \( Y_{t,m} \) where the weights are from the distribution of \( M_{t'} \). When \( M_{t'} \) takes finite number of possible values, this becomes a mixture distribution

\[
p(Y_{t,M_t'} = y) = \sum_m \pi_m p(Y_{t,m} = y), \quad \pi_m = P(M_{t'} = m).
\]

When \( M_{t'} \) takes continuous values, the resulting distribution is sometimes called an infinite mixture

\[
p(Y_{t,M_t'} = y) = \int p(Y_{t,m} = y) p(M_{t'} = m) dm.
\]