1 Introduction to Mediation

Mediation is the process through which a treatment A causes an outcome Y through an intermediate variable M. Consider an example where researchers are interested in whether smoking affects medical expenditures. Smoking is a binary treatment A and medical expenditures is the outcome of interest Y. Researchers may not necessarily be interested in only the effect of smoking on medical expenditures, but also in the effect of smoking on medical expenditures through the pathway of overall perceived health (M). Perhaps smoking itself does not directly cause higher medical expenditures, but rather smoking indirectly causes higher medical expenditures because it decreases patients' overall perceived health. In this example, overall perceived health is considered the mediator. A mediator M is a variable that occurs in time between the treatment A and outcome Y. It is on the causal pathway between the treatment and outcome, and hence may be affected by the treatment and may also affect the outcome itself.

Causal mediation analysis seeks to decompose the total causal effect of the treatment on the outcome into both a direct effect and an indirect effect. The direct effect is the effect of treatment on the outcome without any presence of a mediator, while the indirect effect is the effect of the treatment on the outcome operating through a mediator. There can also consist of a set of confounders. Confounders are variables that influence the treatment and/or outcome. In our example of the effect of smoking on medical expenditures, confounders may include age, income, and education level. When all confounders are measured and controlled for, this allows us to identify the causal effects of interest. Thus, an important statistical challenge is to quantify the direct and indirect effects while properly accounting for the existence of all confounders.

Maybe more about how mediation analysis has been done in literature? Identification and estimation of direct/indirect effects in literature? The goals of our model

2 Potential Outcomes Framework

We will use the potential outcomes framework for causal mediation analysis. For individuals $i = 1, ..., n$ and treatment $A_i \in \{0, 1\}$, define the potential outcome $M_i(a)$ as the value of the mediator that would have been observed had the individual received treatment a. Note that for each individual, only one of $M_i(0)$ or $M_i(1)$ is actually observed. For treated individuals ($A_i = 1$), $M_i(0)$ can be thought of as the counterfactual. That is, the value of the mediator that would have been observed had the individual been untreated instead. Similarly, the potential outcome $Y_i(a, m)$ is the value of the outcome that would have been observed had the individual received treatment a and had a mediator at level m. For example, $Y_i(0, M_i(1))$ is the value of the outcome that would have been observed if the individual was not treated and had a value of the mediator at the same level they would have had if they were treated. Hence, it is only possible to observe $M_i(A_i)$ and $Y(A_i, M_i(A_i))$.

We can use these potential outcomes to define causal estimates of interest. In causal mediation analysis, we are particularly interested in estimating the natural direct and natural indirect effects. The natural direct effect is defined as

$$\zeta(a) = E[Y_i\{1, M_i(a)\} - Y_i\{0, M_i(a)\}] \tag{1}$$

and the natural indirect effect is defined as

$$\delta(a) = E[Y_i\{a, M_i(1)\} - Y_i\{a, M_i(0)\}]. \tag{2}$$

The natural direct effect isolates the effect of the treatment while keeping the potential mediator fixed and can be interpreted as the effect that the treatment A_i has directly on the outcome Y_i. On the other hand, the natural indirect effect isolates the effect of the potential mediator in response to different treatment values while keeping the treatment fixed and can be interpreted as the effect that the treatment A_i has indirectly on the outcome Y_i through the mediator M_i. The total effect of the treatment on the outcome is a sum of the direct and indirect effects. The total effect can be defined as

$$\tau = \zeta(0) + \delta(1) = \zeta(1) + \delta(0) = E[Y_i\{1, M_i(1)\} - Y_i\{0, M_i(0)\}]. \tag{3}$$

3 Assumptions?

4 Bayesian Additive Regression Trees

Consider an unknown function f that predicts an output Y_i using a vector of inputs $\boldsymbol{x_i}$

$$Y_i = f(\boldsymbol{x_i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{4}$$

We can model $f(\boldsymbol{x_i})$ by a sum of m regression trees $f(\boldsymbol{x_i}) \approx r(\boldsymbol{x_i}) = \sum_{j=1}^{m} g(\boldsymbol{x_i}; T_j, M_j)$. T_j is a binary decision tree consisting of a set of interior node decision rules as well as a set of terminal nodes. The decision rules are binary splits of the predictor space. $M_j = \{\mu_{j1}, \mu_{j2}, ...\mu_{jb_j}\}$ is a set of parameter values associated with each of the b_j terminal nodes of tree T_j. Each x_i is associated with a single terminal node of T_j and is then assigned the μ_{jk} value associated with that terminal node. Hence, for a given tree T_j and terminal node parameters M_j, $g(\boldsymbol{x_i}, T_j, M_j)$ denotes the function that assigns a $\mu_{jk} \in M_j$ to $\boldsymbol{x_i}$. Thus (4) can be approximated by the sum-of-trees model

$$Y = \sum_{i=1}^{m} g(\boldsymbol{x_i}; T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{5}$$

Under (5), $E(Y_i|\boldsymbol{x_i})$ equals the sum of all the terminal node u_{jk}'s assigned to $\boldsymbol{x_i}$ by the $g(\boldsymbol{x_i}; T_j M_j)$'s.

It is also necessary to specify a prior over all parameters of the sum-of-trees model, i.e., (T_j, M_j) for all j. This prior should regularize the fit by keeping individual tree effects from being disproportionately influential. The prior consists of two components: a prior for each tree T_j and a prior on the terminal nodes $M_j|T_j$ where $p(T_j, M_j) = p(M_j|T_j) \times p(T_j)$. We can write $r \sim \text{BART}(p(T_j), p(M_j|T_j), m)$. That is, r has a BART prior with m trees, a tree prior $p(T_j)$, and a terminal node prior $p(M_j|T_j)$.

The prior $p(T_j)$ is specified by three aspects: the probability that a node is interior, the distribution of the splitting variable assignments at each interior node, and the distribution of the splitting rule assignment in each interior node conditional on the splitting variable. The probability that a node at depth d is interior is

$$\alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \beta \in [0, \infty) \tag{6}$$

with $\alpha = 0.95$ and $\beta = 2$ being the default that favors small trees. The distribution of the splitting variable assignments at each interior node and the distribution of the splitting rule assignment in each interior node conditional on the splitting variable are both given a uniform prior.

For the prior on the terminal nodes, note that $p(M_j|T_j) = \prod_{k=1}^{b_j}[\mu_{jk}|T_j]$. First, shift and rescale Y so that the observed transformed y values range from $y_{min} = -0.5$ to $y_{max} = 0.5$. Then, we use the prior

$$\mu_{jk}|T_j \sim N(0, \sigma_\mu^2) \quad \text{where } \sigma_\mu = \frac{0.5}{k\sqrt{m}} \tag{7}$$

for a suitable value of k with default $k = 2$. Note that this prior shrinks the terminal node values μ_{jk} towards zero and applies greater shrinkage as the number of trees m is increased, ensuring that each tree is a weak learner in the ensemble of trees.

5 BART for Mediation

We will use BART in modeling the mediator given the treatment and covariates as well as the outcome given the treatment, covariates, and mediator. That is, we model components of $[M_i|A_i, \boldsymbol{X_i}]$ and $[Y_i|M_i, A_i, \boldsymbol{X_i}]$ using BART. These models are discussed in the context of binary, ordinal, and continuous mediators. BCFs and BCMFs, propensity score/clever covariates

5.1 Binary Mediator

For a binary mediator, we specify the outcome and mediator models as follows.

$$Y_i(a,m) = \mu(X_i) + az(X_i) + md(X_i) + \epsilon_i \tag{8}$$
$$M_i(a) = \Phi(\mu_m(X_i) + a\tau_m(X_i)) + \varepsilon_i \tag{9}$$

We use BART priors for μ, z, d, μ_m, and τ_m. The model for M_i is a probit regresssion of M_i, where we use the Albert-Chib truncated normal latent variables z_i where

$$z_i \mid M_i, \mu_m, \tau, A, X_i \sim \begin{cases} N\{\mu_m(X_i) + a\tau_m(X_i), 1\}\, \mathbb{I}(-\infty, 0) & \text{if } M_i(a) = 0 \\ N\{\mu_m(X_i) + a\tau_m(X_i), 1\}\, \mathbb{I}(0, \infty) & \text{if } M_i(a) = 1 \end{cases}$$

so that we can use the continuous outcome BART model of the previous section.

We can compute the natural direct and indirect effects from (1) and (2) as

$$\begin{aligned} \zeta(x) &= E[\mu(X_i) + z(X_i) + M_i(a)d(X_i) - \mu(X_i) - M_i(a)d(X_i) \mid X_i = x] && (10)\\ &= z(x) && (11) \end{aligned}$$

$$\begin{aligned} \delta(x) &= E[\mu(X_i) + az(X_i) + M_i(1)d(X_i) - \mu(X_i) - az(X_i) - M_i(0)d(X_i) \mid X_i = x] && (12)\\ &= d(x)[\Phi(\mu_m(x) + \tau_m(x)) - \Phi(\mu_m(x))]. && (13) \end{aligned}$$

Therefore, $z(x)$ is exactly the direct effect while the indirect effect is the product in (13).

5.2 Ordinal Mediator

5.3 Continuous Mediator

With a continuous mediator, the outcome and mediator models are specified as follows.

$$Y_i(a,m) = \mu(X_i) + az(X_i) + md(X_i) + \epsilon_i \tag{14}$$
$$M_i(a) = \mu_m(X_i) + a\tau_m(X_i) + \varepsilon_i \tag{15}$$

where we again use BART priors for μ, z, d, μ_m, and τ_m.

The natural direct and indirect effects are

$$\begin{aligned} \zeta(x) &= E[\mu(X_i) + z(X_i) + M_i(a)d(X_i) - \mu(X_i) - M_i(a)d(X_i) \mid X_i = x] && (16)\\ &= z(x) && (17) \end{aligned}$$

$$\begin{aligned} \delta(x) &= E[\mu(X_i) + az(X_i) + M_i(1)d(X_i) - \mu(X_i) - az(X_i) - M_i(0)d(X_i) \mid X_i = x] && (18)\\ &= d(x)\tau_m(x). && (19) \end{aligned}$$

6 Medical Expenditures Panel Survey Data

As a demonstration of our model, we consider the medical expenditures panel survey dataset. The goal of this survey was to examine the effect of smoking on medical expenditures. For our purposes, we want to examine both the direct effect of smoking on medical expenditures as well as the indirect effect of smoking on medical expenditures through the mediator of overall perceived health. More about the subet of data used? Therefore, the outcome Y is the natural logarithm of medical expenditures, the treatment A is whether or not an individual smokes (0: non-smoke, 1: smoke), and M is an ordinal measure of overall perceived health (1: excellent, 2: very good, 3: good, 4: fair, 5: poor). Our model also includes the following patient attributes as confounders:

- `age`: age in years at the time of the survey
- `race_white`: white or non-white
- `income`:
- `bmi`: bmi at the time of the survey
- `education_level`:
- `poverty_level`: