PSC 504: Causal Mechanisms

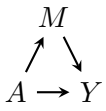
Matthew Blackwell

4/11/2013

Causal Mechanisms

Setup

- In the social sciences, we have theories and these theories tell us we should see certain causal effects, but they also tell us **how** those causes should impact the outcomes. One theory might imply one "causal path" and another theory might imply a completely different causal path. How do we adjudicate between these theories when they predict the same overall effect?
- Trying to investigate the various causal pathways is a branch of causal inference we call causal mechanisms. A causal mechanism is the set of paths through which the effect of the treatment on the outcome flows.
- An example from Imai, Keele, and Yamamoto (2010) is that of media framing. A classic study randomly assigned participants to watch one of two news stories about a KKK rally: one that emphasized free speech concerns and one that emphasized potential violence. The authors of the study thought that the effect of the frame on tolerance for the KKK would be mediated by people's views on intermediate views on the importance of free speech and public order. That is, we might have something like this:



- As usual, we have our treatment variable, A_i and our outcome variable Y_i, but now we have an intermediate, post-treatment variable, M_i, which we call a mediator. Because this is a post-treatment variable, it has potential responses, $M_i(a)$, which is the value that the mediator takes when the treatment is a. The outcome has joint potential outcomes: $Y_i(a, m)$. This is the value that the outcome takes when the treatment has value a and the mediator takes the value m.
- We have to make a consistency assumption, as usual, to connect the potential outcomes to the observed outcomes. In this case, we have to make a consistency assumption for both the mediator and the outcome. Specifically, we will assume that that $M_i = M_i(A_i)$ and that $Y_i = Y_i(A_i, M_i(A_i))$. Thus, the observed mediator is the potential outcome for the mediator under the observed treatment. Note that we can also write components of the potential outcomes: $Y_i(a) = Y_i(a, M_i(a))$. The potential outcome under a is the outcome we would see under a and the value that the mediator would take under a.

Estimands

- There are a couple of different quantities we might want to estimate here. There is the typical individual causal effect, which we will call the **total causal effect**:

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- The total causal effect allows the effect of the treatment "propogate" through all causal pathways. Thus, the mediator notation here is redundant.

- The DAG that we drew above, though, implies that we might be able to think about "direct" and "indirect" effects of the treatment. "Indirect" here is the part of the effect of treatment that "flows through" the mediator and the direct effect is the part of the effect that does not flow through the mediator. These definitions are a little imprecise, so we will specify exactly what they mean.

- One estimand is the so-called "natural" or "pure" indirect effect:

$$\delta_i(a) = Y_i(a, M_i(1)) - Y_i(a, M_i(0))$$

- Note what is happening here. We are fixing the value of the treatment and seeing how the effect of A_i on M_i changes the outcome. Note that one of the two quantities will not be observeable ever. Take $Y_i(1, M_i(0))$. This is the value Y_i would take if the a unit were treated, but we set the mediator to value it would take under control. Obviously, we can't simultaneously see how someone responds to treatment (for the outcome) and how they respond to control (for the mediator). This is different than the "fundamental problem of causal inference," where we only observe treatment or control for a given unit, so we can only observe one of the two possible potential outcomes. Here, it is impossible (without further, strong assumptions) to even observe these "counterfactual" potential outcomes. Rubin has made strong claims that even investigating quantities like these is "unscientific."

- Also, note that we are also assuming that the **way** we affect the mediator does not matter. That is, if $M_i(1) = M_i(0) = m$, then $Y_i(a, M_i(1)) = Y_i(a, M_i(0)) = Y_i(a, m)$. So it doesn't matter if the treatment sets the mediator or we set the mediator by intervention.

- Imai, Keele, and Yamamoto (2010) focus on the average of these indirect effects, which they call the average causal mediation effect (ACME):

$$\bar{\delta}(a) = E[\delta_i(a)] = E[Y_i(a, M_i(1)) - Y_i(a, M_i(0))]$$

- We can also define the natural/pure direct effect (PDE) of the treatment:

$$\zeta_i(a) = Y_i(1, M_i(a)) - Y_i(0, M_i(a))$$

- Thus, the pure direct effect is the effect of moving from control to treatment while holding the mediator fixed at the value it would have under treatment status a.

- Why might we care about a quantity like this? The canonical example involves smoking as the treatment and tar as a mediator, with lung cancer as an outcome. We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$, and that, overall, smoking increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$. But we may want to know about what would happen if we created a cigarette without any tar in it.

- Note that the total causal effect and the pure indirect and direct causal effects are related:

$$\tau_i = \delta_i(a) + \zeta_i(1-a)$$

- Thus, we know that the ATE, $\bar{\tau} = E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$, must be the sum of the average indirect and direct effects:

$$\bar{\tau} = \bar{\delta}(a) + \bar{\zeta}(1-a)$$

- The fact that we can decompose the total effect of treatment into the sum of a direct and indirect effect if very important to social science researchers.

- One more estimand that we might be interested in is similar to our investigations into panel data and fixed effects. That is the **controlled direct effect** (CDE):

$$Y_i(1, m) - Y_i(0, m)$$

- In general, this effect will be different than the PDE.

Identification

- We know what assumptions identify the ATE and how to estimate it with data, but can we estimate this new quantity, the ACME? Yes, we can, but we need more assumptions than usual.

- Imai, Keele, and Yamamoto (2010) use an assumption they call **sequential ignorability**, which is actually the same name as a different assumption in the context of time-varying treatments. The assumption has two parts. First, the treatment must be ignorable with respect to the mediator and the outcome:

$$\{Y_i(a', m), M_i(a)\} \perp\!\!\!\perp A_i | X_i = x$$

- Here, this has to hold for all a, a' and all m. This assumption could be satisfied with a randomly assigned treatment.

- The next part of the assumption is that the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(a', m) \perp\!\!\!\perp M_i(a) | A_i = a, X_i = x$$

- Here, again, this holds for all values of a, a'. Note that we have to believe ignorability in certain cross-world comparisons:

$$Y_i(1, m) \perp\!\!\!\perp M_i(0) | A_i = 0, X_i = x$$

- This says that the mediator value under control is independent of what happens to the outcome under treatment. This is a strange ignorability assumption and has a strange property: randomization of the treatment and mediator does not imply this assumption holds. This is because we need to make assumptions about the potential values of the mediator—how the mediator responds to the treatment. If we set the mediator ourselves via randomization, we would lose that crucial information.

- Note also, that the mediator ignorability must hold only on pre-treatment covariates, not post-treatment variables (that is, other potential mediators).

- Under these two assumptions, we can write the ACME as a function of the observed data. It's easy to write this out when the mediator has J categories:

$$\begin{aligned}\bar{\delta}(a) = \sum_{m=0}^{J-1} & E[Y_i|M_i = m, A_i = a, X_i] \\ & \cdot \{\Pr[M_i = m|A_i = 1, X_i] - \Pr[M_i = m|A_i = 0, X_i]\}\end{aligned}$$

- What does this look like? It's the effect of the treatment within a level of the mediator multiplied by the effect of the treatment on the probability of seeing that value of the mediator. It's a little more clear with a binary mediator:

$$\begin{aligned}\bar{\delta}(a) =& \{\Pr[M_i = m|A_i = 1, X_i] - \Pr[M_i = m|A_i = 0, X_i]\} \\ & \cdot \{E[Y_i|M_i = 1, A_i = a, X_i] - E[Y_i|M_i = 0, A_i = a, X_i]\end{aligned}$$

- The first term here is the effect of the treatment on the mediator and the second term is the effect of the mediator on the outcome, conditional on the treatment.

Linear Structural Equation Models

- Let's say that we have a linear, structural model for all variables:

$$\begin{aligned}M_i(a) &= \alpha_0 + \alpha_1 a + \eta_i \\ Y_i(a, m) &= \beta_0 + \beta_1 a + \beta_2 m + \varepsilon_i\end{aligned}$$

- It's clear that we can write the total effect of the treatment in the following way:

$$\begin{aligned}Y_i(1, Z_i(1)) - Y_i(0, Z_i(0)) =& \beta_0 + \beta_1 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\ & - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\ =& \beta_1 + \beta_2 \cdot \alpha_1\end{aligned}$$

- What about the indirect effect:

$$\begin{aligned}Y_i(0, Z_i(1)) - Y_i(0, Z_i(0)) =& \beta_0 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\ & - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\ =& \beta_2 \cdot \alpha_1\end{aligned}$$

- If we think that we can identify these regressions, say because we randomly assigned the treatment and the mediator, then we can estimate the total effect from a regression of Y_i on A_i and X_i, then we can estimate the direct effect (β_1) from a regression of Y_i on A_i, M_i, and X_i. Then we can take the difference between these coefficients and get the indirect effect of the treatment.

- Note, though, that there is an implicit assumption here of no interactions between the indirect effect and the treatment status. That is,

$$\bar{\delta}(1) = \bar{\delta}(0)$$

- We could incorporate an interaction into the model here to allow for the indirect effect to vary.
- Under sequential ignorability, we can estimate α_1 from a regression of M_i on A_i and then estimate β_2 from a regression of Y_i on A_i) and M_i. Then, our estimate of the ACME is simply the product of these estimates: $\hat{\delta} = \hat{\alpha}_1\hat{\beta}_2$). The variance of this estimator can be written:

$$V[\hat{\delta}] = \alpha_1^2 V[\hat{\beta}_2] + \beta_2^2 V[\hat{\alpha}_1] + V[\hat{\beta}_2]V[\hat{\alpha}_1]$$

Nonparametric Estimation

- The above estimators assume that treatment effects are constant across units and that everything is linear. These may not be valid assumptions. Instead, we may want to estimate the effects with non-parametric estimators.
- If the number of categories in the mediator is small, then we can fill in the conditional expectations above with their sample counterparts.
- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator. To get the standard errors, we can use bootstrapping.
- What if the mediator is continuous? Things get tricky. This is because we have to integrate over the distribution of the mediators to get the ACME:

$$\bar{\delta}(a) = \int\int E[Y_i|M_i = m, A_i = a, X_i = x] \{dF_{M_i|T_i=1,X_i=x}(m) - dF_{M_i|T_i=0,X_i=x}(m)\} dF_{X_i}(x)$$

- Obviously, this is a much harder problem. In this case, we actually can use Monte Carlo simulation to take the integral.