Identifying causal mechanisms in experiments (primarily) based on IPW

Martin Huber

University of St. Gallen, School of Economics and Political Science

Field experiments in policy evaluation

Oct 18-19 2012, Nuremberg

Outline of the talk

In a nutshell:

- Identification of (natural/pure) direct and indirect effects based on inverse probability weighting by the treatment propensity score
  - Random assignment of binary treatment
  - conditional exogeneity of the mediator given the treatment and observed covariates
  - Two scenarios: mediator exogeneity holds conditional on (i) pre-treatment or (ii) post-treatment covariates

Outline:

- Introduction
- Assumptions and identification
- Empirical application
- Conclusion

Introduction

Identifying causal mechanisms in experiments:

- Randomized experiments are widely used in social sciences and often regarded to be the gold standard of causal inference.
- In many economic problems, not only the (total) effect of an intervention (such as the ATE) appears relevant, but also the causal mechanisms (through mediators) through which it operates.
- However, random treatment assignment does not imply the randomness of (post-treatment) mediators (see for instance Rosenbaum (1984) and Robins and Greenland (1992)), which are themselves intermediate outcomes.
- This requires us to control for confounders of the mediator (in a way that does not break treatment randomization).

Introduction

Main contribution:

- Nonparametric identification of direct and indirect effects mainly based on inverse probability weighting (IPW) (see Horvitz and Thompson (1952)), assuming that the confounders of the mediator are observed.
- Observations are weighted by the inverse of their conditional propensity to be (non-)treated given the mediator and the observed covariates.
- The identification results are attractive from a practitioner's point of view, because they only involve the estimation of a propensity score model instead of conditional expectations of outcomes and conditional densities of mediators.
- The identification results depend on whether the covariates are themselves a function of the treatment or not.
- If they are, identification of the (total) indirect effect requires additional functional form assumptions, see Robins (2003), Avin, Shpitser, and Pearl (2005), and Imai and Yamamoto (2011).

Further literature on non- and semiparametric identification based on observed covariates:

- In statistics/social sciences: Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), van der Weele (2009), Imai, Keele, and Yamamoto (2010), Albert and Nelson (2011), and Imai and Yamamoto (2011), among others.

- In economics, comparably few: Simonsen and Skipper (2006), Flores and Flores-Lagunes (2009) - both only use pre-treatment covariates to control for mediator endogeneity, which might be implausible given that the mediator is a post-treatment variable.

Assumptions and identification

Notation:

- Y: outcome of interest (discrete or continuous)
- D: binary treatment (randomly assigned)
- M: mediator (discrete or continuous)
- X: observed covariates (multidimensional, discrete and/or continuous)
- $M(d)$: potential mediator for $d \in \{0,1\}$
- $Y(d, M(d'))$: potential outcome for $d, d' \in \{0,1\}$
- i.i.d. sampling

Assumptions and identification

Average direct and indirect effects:

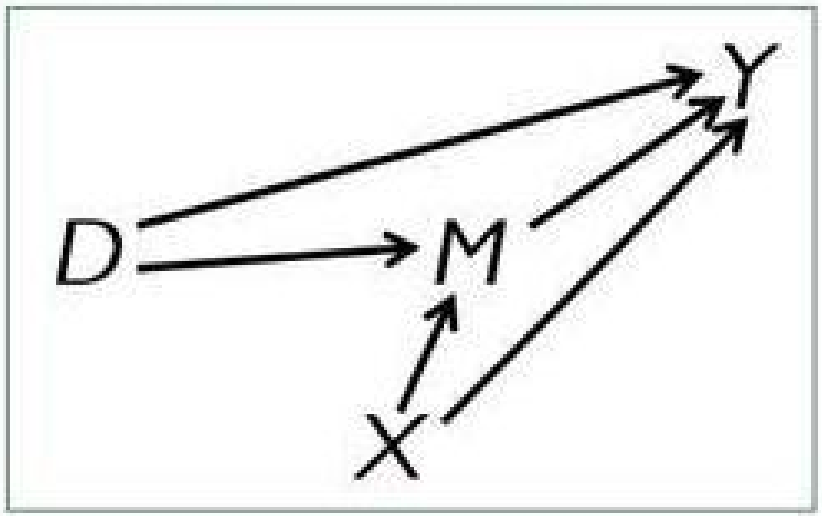
$$\begin{aligned} \theta(d) &= E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}, \\ \delta(d) &= E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \end{aligned}$$

Identification issues:

- $Y = D \cdot Y(1, M(1)) + (1 - D) \cdot Y(0, M(0))$
- $M = D \cdot M(1) + (1 - D) \cdot M(0)$
- Either $Y(1, M(1)), M(1)$ or $Y(0, M(0)), M(0)$ is observed for a particular observation.
- $Y(1, M(0))$ and $Y(0, M(1))$ are never observed.

First scenario considered:

Figure 1: X is not a function of D



Assumptions and identification

Assumption 1 (random treatment assignment):

$\{Y(d', m), M(d), X\} \perp D$ for all $d', d \in \{0, 1\}$ and m in the support of M.

- D is independent of X and of any unobservable factors jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand.
- Assumption 1 also implies that $\{Y(d', m), M(d)\}$ is independent of D given X.

Assumption 2 (conditional independence of the mediator):

(a) $Y(d', m) \perp M | D = d, X = x$ for all $d', d \in \{0, 1\}$ and m, x in the support of M, X,

(b) $\Pr(D = d | M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

- Conditional on D and X, the effect of the mediator on the outcome is assumed to be unconfounded.
- Assumption 2(b) is a common support restriction requiring that the treatment propensity score is larger than zero in either treatment state, which implies that $\Pr(M = m | D = d, X = x) > 0$.

Proposition 1:
Under Assumptions 1 and 2, the average direct effect is identified by

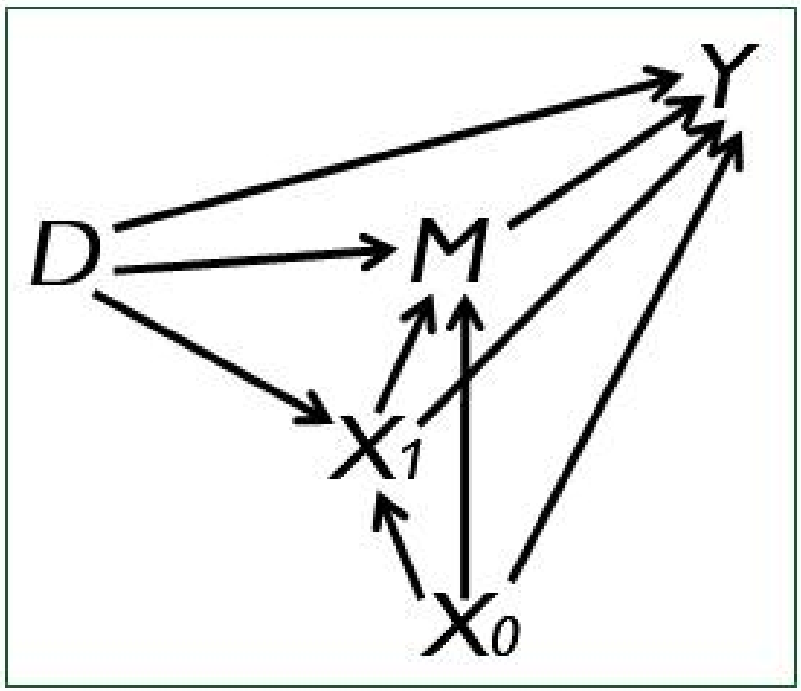
$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)}\right].$$

Proposition 2:
Under Assumptions 1 and 2, the average indirect effect is identified by

$$\delta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1)}\right)\right].$$

Second scenario considered:

Figure 2: X is (partially) a function of D



Assumptions and identification

Modification in notation:

- $X(d)$: Potential state of the covariates for $d \in \{0, 1\}$.
- $M(d) = M(d, X(d))$
- $Y(d, M(d)) = Y(d, M(d, X(d)), X(d))$

Direct effect:

$$\theta(d) = E[Y(1, M(d, X(d)), X(d)) - Y(0, M(d, X(d)), X(d))],$$

Total indirect effect (all effects via M which either come from D directly or "take a devious route" through X):

$$\delta^t(d) = E[Y(d, M(1, X(1)), X(d)) - Y(d, M(0, X(0)), X(d))].$$

Partial indirect effect (only identifies the effect through M directly coming from D, but not going through X):

$$\delta^p(d) = E[Y(d, M(1, X(d)), X(d)) - Y(d, M(0, X(d)), X(d))].$$

Assumptions and identification

Assumption 3 (random treatment assignment and conditional independence):

(a) $\{Y(d'', m, x'), M(d', x), X_0, X_1(d)\} \perp D$ for all $d'', d', d \in \{0, 1\}$ and m, x, x' in the support of M, X,

(b) $\{Y(d'', m, x''), M(d', x')\} \perp D | X = x$ for all $d', d \in \{0, 1\}$ and m, x'', x', x in the support of M, X.

- Assumption 3(a) no longer imposes independence of D and X, but merely random assignment of the treatment.
- Assumption 3(b) is new and explicitly states that unconfoundedness of the treatment effects on the mediator and outcome must also hold when conditioning on X.
- This implies that there are no unobserved confounders which jointly affect X on the one hand and M and/or Y on the other hand.
- Conditioning on post-treatment variables changes the distribution of pre-treatment variables across treatments (which initially were balanced by randomization) $\rightarrow$ need to control for pre-treatment covariates related with the post-treatment confounders of the mediator and with the outcome/mediator directly.

Assumptions and identification

Assumption 4 (conditional independence of the potential mediator state):

(a) $Y(d'', m, X(d')) \perp M | D = d, X = x$ for all $d'', d', d \in \{0, 1\}$ and m, x in the support of M, X,

(b) $\Pr(D = d | M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

- Assumption 4(a) is equivalent to Assumption 2(a), but now accounts for our modified notation.
- The common support restriction 4(a) is exactly the same as before.

Proposition 3:
Under Assumptions 3 and 4, the average direct effect is identified by

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)}\right].$$

Proposition 4:
Under Assumptions 3 and 4, the average partial indirect effect is identified by

$$\delta^p(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)} - \frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)} \cdot \frac{\Pr(D=d|X)}{\Pr(D=d)}\right].$$

Assumptions and identification

Assumption 5 (functional form restriction w.r.t. potential mediators):

For all m_d, x_d in the support of $M(d), X(d)$, write $E[Y(d, M(d, X(d)), X(d))|M(d, X(d)) = m_d, X(d) = x_d] = \mu_{d,x_d}(m_d)$, i.e., write the mean potential outcome for $D = d$ as a function of m_d.

It is assumed that

(a) for all m_{1-d}, x_d in the support of $M(1-d), X(d)$, it holds that $\mu_{d,x_d}(m_{1-d}) = E[Y(d, M(1-d, X(1-d)), X(d))|M(1-d, X(1-d)) = m_{1-d}, X(d) = x_d]$,

(b) $\mu_{d,x_d}(m_{1-d}) = \mu_{d,x_d}(E(m_{1-d}))$.

Assumptions and identification

- Assumption 5(a) states that one can predict $E[Y(d, M(1-d, X(1-d)), X(d))|M(1-d, X(1-d)) = m_{1-d}, X(d) = x_d]$ for any m_{1-d} and x_d based on the regression function μ_{d,x_d}. This implies that the interaction effect between D and M is the same for $M(1)$ and $M(0)$.
- $M(1-d, X(1-d))$ is not known for units with $D = d$ (on which the identification of μ_{d,x_d} is based upon), but $E[Y(d, M(1-d, X(1-d)), X(d))] = E[Y(d, E[M(1-d, X(1-d))], X(d))]$ by Assumption 5(b) and $E[M(1-d, X(1-d))]$ is observed for $D = 1-d$.
- Assumption 5 is not innocuous. Firstly, Assumption 5(a) requires a correctly specified model for the prediction across mediator states. Secondly, Assumption 5(b) restricts μ to be linear in M such that predicting based on $E(m_{1-d})$ is asymptotically equivalent to using m_{1-d}.

Assumptions and identification

Proposition 5:
Under Assumptions 3, 4, and 5, the average total indirect effect is identified by

$$\delta^t(d) = E\left[\frac{\{Y - \mu_{d,x}(E[M|D = 1 - d])\} \cdot I\{D = d\}}{\Pr(D = d)}\right].$$

Empirical application

- Experimental evaluation of Job Corps, an educational program targeting young individuals (aged 16-24 years) from low-income households.
- D: Assignment to the program (random)
- Y: Indicator of very good health 2.5 years after randomization
- M: Employment (binary) 1 to 1.5 years after randomization
- X: health, labor market history, and socio-economic status shortly prior to the mediator and at treatment assignment
- Evaluation data consist of 4,352 females and 5,673 males.

Table 1: Effects on the incidence of very good general health after 2.5 years

	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}^t(1)$	$\hat{\delta}^t(0)$	$\hat{\delta}^p(1)$	$\hat{\delta}^p(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
					Females				
Effect	0.028	0.031	0.023	-0.000	0.000	-0.000	0.000	0.005	-0.003
S.E.	0.014	0.018	0.015	0.001	0.001	0.000	0.001	0.006	0.011
p-value	0.045	0.078	0.123	0.737	0.783	0.784	0.891	0.405	0.813
					Males				
Effect	0.022	-0.003	0.003	-0.000	-0.000	0.000	0.000	0.019	0.025
S.E.	0.013	0.019	0.014	0.000	0.000	0.000	0.000	0.007	0.013
p-value	0.099	0.861	0.845	0.584	0.782	0.540	0.677	0.004	0.060

Note: Standard errors (S.E.) are estimated based on 1999 bootstrap draws.

Conclusion

- This paper discusses the identification of direct and indirect effects in randomized experiments with a binary treatment (mainly) based on inverse probability weighting (IPW) using the treatment propensity score.
- Identification relies on the assumption of conditional exogeneity of the mediator given observed covariates and the treatment.
- Identification results are discussed for two sets of assumptions: Mediator exogeneity (i) given covariates which are not influenced by the treatment (with the leading case being pre-treatment variables) and (ii) given covariates which are themselves a function of the treatment.
- An empirical application to the Job Corps experimental study is also provided.