The evaluation of causal effects and mechanisms in empirical economics

Martin Huber

University of St. Gallen

1 Introduction

My research aims at improving and applying statistical and econometric methods for the analysis of causal effects in empirical economics (i.e. based on real-world data), in particular in labor, education, and health economics. Amongst other research questions, I have used such methods to investigate for instance the effects of education and training programs on individual earnings and employment, the impact of welfare dependency on health and mental wellbeing, and the influence of preschool attendance on the development of cognitive skills of minority children.

More specifically, my studies focus on the development and application of so-called micro-econometric estimators and tests as often applied to labor market or health data, with the goal to obtain ever more appropriate statistical tools for the measurement of causal effects. "More appropriate" might imply that a new method relies on less statistical assumptions that restrict human behavior in a particular way, or that such assumptions can (if they must be maintained) at least be empirically tested by checking whether they contradict behavioral patterns seen in the data.

Two major challenges in causal analysis are so-called endogeneity and attrition problems. Endogeneity implies that the causal effect of a variable of interest cannot be easily measured because it interacts (and from the researchers perspective awkwardly so) with other, frequently unobserved variables (e.g. education is associated with unobserved ability and both affect wage, so that the wage effect of education alone is hard to identify). Attrition implies that some individuals who are initially observed in the data drop out in later periods (e.g. due to reluctance to participate in a follow-up survey). Both problems generally jeopardize the causal interpretability of statistical/econometric results. The majority of my work is therefore dedicated to the development of (i) methods that can deal with endogeneity and/or attrition issues or (ii) tests that allow detecting such problems in the data.

I subsequently discuss in more detail one topic among my research interests, namely the assessment of causal mechanisms in empirical economics, and how the mentioned endogeneity issues may be addressed in this context.

2 A conceptual framework for causal effects and mechanisms

In empirical economics we typically aim at evaluating the causal effect of an explanatory variable (also called "treatment") on an outcome of interest. Most studies focus on the identification of the "total" causal effect. That is, they do not consider the possibility that the latter may come from distinct causal mechanisms due to intermediate variables (hereafter "mediators", as the analysis of their impact is often called "mediation analysis") that lie on the causal path between the treatment and the outcome of interest. In the presence of one or more mediators, the total effect can be decomposed into a direct effect (or net effect) of the treatment on the outcome of interest and an indirect effect which operates through the mediators. Such a decomposition of causal mechanisms often provides a better understanding of the economic problem than the total effect alone and may be crucial for deriving meaningful policy conclusions.

When for instance evaluating the employment effects of a training program for job seekers, we face the problem that the latter could induce further participation in programs which themselves affect employment. Disentangling the causal mechanisms tells us whether the initial program is effective per se (net of further participation) or only jointly with further interventions, an information required for the optimal design of labor market policies. A further example is the decomposition of the effectiveness of the entire job counseling process provided by local employment offices on the labor market success (e.g., employment and earnings) of job seekers. One interesting question is whether counseling affects labor market success exclusively through placement into training programs or also has a direct effect related to the counseling style and cooperativeness of the case worker with the job seeker (net of training). This appears important for developing guidelines for an efficient counselling process.

It is useful to introduce some notation to characterize the econometric problems related to the assessment of causal effects and mechanisms. For the sake of simplicity, only a binary treatment with values 1 if "treated" (e.g. training participation) and 0 if "not treated" (no training) and a single mediator are considered. Let Y, M, D denote the outcome, the mediator, and the binary treatment indicator, respectively, that are *actually observed* for all individuals in the researcher's data set. Furthermore, let $Y(1), M(1)$ denote the potential outcome and the potential mediator state an individual would have *hypothetically* realized under treatment, and $Y(0), M(0)$ the hypothetical realisations under non-treatment. Note that depending on the *actually* realized treatment state (either 1 or 0), either $Y(1), M(1)$ or $Y(0), M(0)$ are known for some individual, but never both pairs, as an individual cannot be treated and non-treated at the same time. That is, for the treated, Y, M corresponds to $Y(1), M(1)$, while for the non-treated, Y, M corresponds to $Y(0), M(0)$.

An interesting effect often considered in empirical economics is the average causal effect (ACE) of the treatment on the outcome (e.g. employment or earnings) in some population, e.g. among all jobseekers in Switzerland:

$$\Delta = E[Y(1) - Y(0)],$$

where "E" stands for mean, average, or expectation. That is, the ACE, denoted by Δ, is simply the mean difference in the potential outcomes (such as potential earnings) when participating vs. not participating in a training among the population of Swiss job seekers. As both $Y(1)$ and $Y(0)$ are not observed at the same time for some job seeker, the ACE can, however, only be assessed based on specific statistical assumptions/designs. The crucial condition for evaluating the ACE is that those actually receiving the training ($D = 1$) are comparable in terms of any characteristics affecting both the training probability and the earnings outcome to those not receiving the training ($D = 0$).

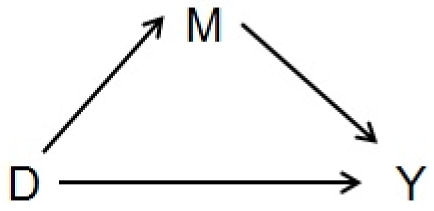
If for instance motivation increases both the chances to obtain the training and the earnings, then the average earnings difference between treated and non-treated is partly driven by differences in motivation and not by differences in training alone, so that the ACE of training cannot be measured. Statistical methods for assessing the ACE therefore aim at making treated and non-treated observations comparable in such characteristics. The most convincing (but not always feasible) method is a randomized experiment: The training is randomly assigned to job seekers, i.e. some receive it by pure chance while others do not. Then, any individual characteristics like motivation should be comparable across treated and non-treated groups, as such factors do not influence the chance to obtain the training. For this reason, also so-called non-experimental methods try to make treated and non-treated individuals comparable in terms of characteristics affecting both the treatment and the outcome, just as the experimental benchmark does by design.

The analysis of causal mechanisms requires disentangling the (total) ACE Δ into a direct and indirect component (through M, e.g. a second training taking place after the first one). We to this end change the (standard) notation for the potential outcome and rewrite it as a function of both the treatment and the intermediate variable M: $Y(1) = Y(1, M(1))$, $Y(0) = Y(0, M(0))$. This modification acknowledges the possibility that part of the causal effect might go directly from the treatment to the outcome (direct effect), while some other part might work through a change in the mediator (indirect effect). It follows that the ACE may be rewritten in the following way:

$$\Delta = E[Y(1) - Y(0)] = E[Y(1, M(1)) - Y(0, M(0))].$$

Figure 1 provides a graphical illustration of our causal framework, in which each arrow represents the causal effect of one variable on another.

Figure 1: Direct and indirect effects from D to Y



To measure the size of the direct effect (or arrow) of D on Y, one would like to keep the mediator constant at a particular value, e.g. $M(0)$ (the potential mediator under non-treatment), in order to “block” the indirect effect (such that only the direct one remains). Therefore, the (average) direct effect (denoted by θ) is identified by

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

i.e., by varying the treatment but keeping the mediator fixed at its potential value for $D = 0$. Conversely, to measure the (average) indirect effect (denoted by δ), one would like to hold the treatment constant, e.g. at $D = 1$, to block the direct effect and only vary the mediator according to its potential values under treatment and non-treatment:

$$\delta = E[Y(1, M(1)) - Y(1, M(0))],$$

i.e., the mediator is shifted from $M(1)$ to $M(0)$ but the treatment is fixed at $D = 1$. Note that the ACE is, of course, the sum of the direct and indirect effects:

$$\begin{aligned}\Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta + \delta,\end{aligned}$$

which follows from adding and subtracting $E[Y(1, M(0))]$. It is obvious that the direct and indirect effects cannot be identified without further assumptions, firstly, because only either $Y(1, M(1))$ or $Y(0, M(0))$ is observed for any individual – the same problem as in the measurement of the ACE – and secondly, because $Y(1, M(0))$ and $Y(0, M(1))$ are not observed for any individual – a problem specific to the evaluation of direct and indirect effects.

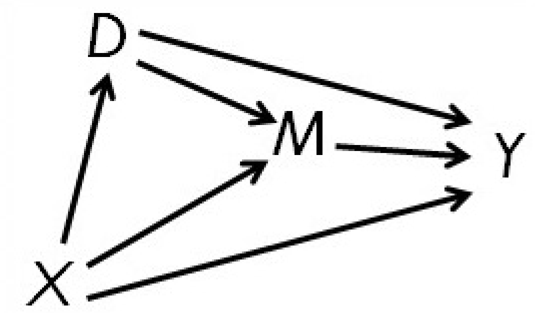
3 Strategies for measuring causal mechanisms

In general, two types of strategies or sets of statistical assumptions can be used to measure the effects of interest despite the problems mentioned in the last section. The first relies on the assumption that we observe all characteristics that jointly affect the treatment and outcome, treatment and mediator, or mediator and outcome in the data set the analysis is based on. This implies that we can make groups in different actual treatment states (e.g. with and without first training) and mediator states (e.g. with and without second training) comparable in terms of these characteristics (e.g. motivation) in order to measure the direct and indirect effects on the outcome (e.g. earnings). Figure 2 illustrates such a set-up (known as “selection on observables”), in which the characteristics to be observed are denoted by X and may have an impact on D, M, and Y.

Given this assumption, it can be shown that the direct and indirect effects may be measured by rather simple formulae involving parameters that are either directly observed in the data (Y, D, M, X) or can be constructed from the data (namely $\Pr(D = 1|X), \Pr(D = 1|M, X)$):

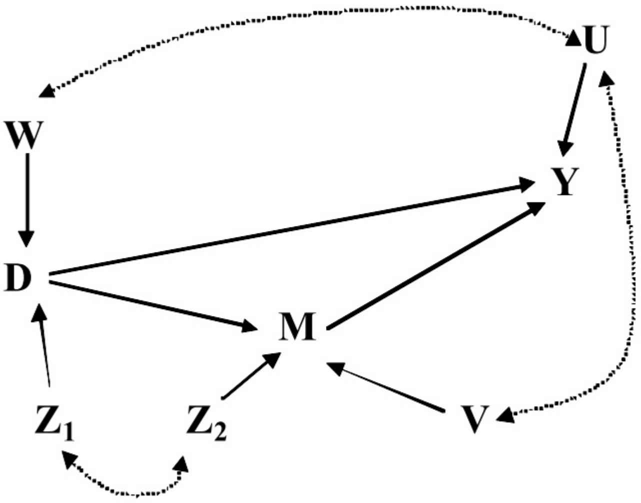
$$\begin{aligned}\theta &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right], \\ \delta &= E\left[\frac{Y \cdot D}{\Pr(D=1|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right],\end{aligned}$$

Figure 2: Direct and indirect effects under “selection on observables”



where $\Pr(D=1|X)$ is the probability to get the treatment for particular values of X (e.g. the likelihood to obtain a training under a low level of motivation) and $\Pr(D=1|M,X)$ is the probability to be treated for particular values of X and M.

Figure 3: Direct and indirect effects with instruments



A second strategy consists of assuming that not all characteristics jointly affecting D, M, Y are observed, but alternatively, that there exists a variable Z_1 that affects D and another variable Z_2 that affects M which satisfy particular properties. A crucial requirement is that Z_1, Z_2 do not have a direct impact on Y, i.e. other than through D and M, respectively. Then, Z_1 and Z_2 are called “instruments” (for D and M, respectively), see the illustration in Figure 3. Note that doubly edged arrows represent potential causal effects in either direction. In such a set up, direct and indirect effects may be measured despite the presence of the assumably unobserved characteristics denoted by U, V, W, which affect each other and D, M, Y (which creates similar problems as a single unobservable characteristic that jointly influences D, M, Y). Depending on the properties of Z_1 and Z_2, one can again derive particular formulae for the effects based on parameters observed/derived in the data. However, it is important to point out that plausible instruments (that do not influence the outcome directly) are not easily obtained in most applications. The credibility of one or the other strategy may therefore differ from one empirical evaluation problem to another.

My current work is based on both strategies and focusses on the development of methods which require a minimum of statistical assumptions (so-called "nonparametric econometrics"), in order to impose as few restrictions as possible on human behavior. Furthermore, I apply these methods to empirical problems in labor and health economics, as for instance (i) to measure the direct effect of an educational program for disadvantaged youths on earnings or subjective health outcomes as well as the indirect effect (via the mediator hours worked) or (ii) to evaluate the direct and indirect effects (via training participation) of the job counselling process in Switzerland, as already discussed before.