Implementation Neutrality and Treatment Evaluation

Stephen F. LeRoy
University of California, Santa Barbara

December 5, 2015

Abstract

Statisticians have proposed formal techniques for evaluation of treatments, often without explicitly specifying how treatments are generated. Under such procedures they run the risk of attributing causation in settings where the implementation neutrality assumption required for causation to be quantitatively meaningful is not satisfied. When treatment assignments are explicitly modeled, as economists recommend, these issues can be formally analyzed. Examples are given.

Statisticians associated with a number of disciplines—medicine, for example—have produced a literature considering how to handle counterfactuals in evaluating the effectiveness of treatments. When randomization of treatments is available, as it usually is in the medical context, the existence of counterfactuals poses no special problems. In some medical and almost all economic contexts, however, one cannot realistically view the assignment of subjects to treatment or lack of treatment as random: the people who are treated differ from those who are not, and ignoring such selection problems may lead to bias. Economists recommend handling this problem by including in their models an explicit specification of the assignment mechanism. Only by so doing is it possible to determine whether a bias exists, and if so how to correct for it.

As many have noted, non-economists resist this approach. They instead propose mechanical algorithms that purportedly make possible diagnosis of causal relations without committing to any particular representation of the assignment mechanism (Pearl [2001], Spirtes, Glymour and Schienes [1993]). Economists—notably Heckman [2001] and elsewhere—have expressed doubts that there is any hope of determining unbiased estimators of causal parameters without committing to an explicit model that includes a characterization of treatment assignment.

Heckman’s concerns are well taken. We demonstrate this in the context of two examples. The first example is set out using econometricians’ analytical framework. It is demonstrated (Section 1) that some causal statements suggested by this model are available only in special cases when conditions for implementation neutrality (defined below and analyzed more fully in LeRoy
are satisfied. Other causal statements, however, do not require implementation neutrality. The analysis is then recast in the framework used in the treatment evaluation literature (Section 2), and we show that this alternative representation implicitly assumes satisfaction of an implementation neutrality property that is generally not satisfied, resulting in the apparent validity of causal statements involving treatment effects that in fact are unjustified. In Section 3 the analysis of Section 1 is extended to statements other than the simplest causal relations.

A second example, discussed in Section 4, involves an evaluation of instrumental variables estimators in settings where the Gauss-Markov assumption underlying ordinary least squares estimators is violated.

1 The Econometric Approach

An econometric model intended to generate causal statements requires explicit specification of variables and a labeling of each variable as external or internal. When there exists a unique equilibrium, as occurs here, the model consists of a vector function mapping external variables to internal variables.

In the case of our example the external variables consist of three random variables: $R$, $U$, and $V$. $R$ is an observable binary random variable that is interpreted as an agent’s race. For simplicity $R$ is assumed to take on value 0 or 1 with probabilities 1/2, 1/2. $U$ and $V$ are unobservable real-valued random variables, distributed uniformly on the unit interval. The internal variables consist of the binary-valued treatment variable $T$, which takes on value 1 if the agent is treated and 0 if not, and the real-valued outcome variable $Y$. The external variables are independently distributed.

The model consists of the following equations:

$$ Y = \alpha_Y T + \beta_Y R + U $$

$$ T = \begin{cases} 
1 & \text{if } V \geq \varphi, \ R = 0 \\
0 & \text{if } V < \varphi, \ R = 0 \\
1 & \text{if } V \geq \varphi + \lambda, \ R = 1 \\
0 & \text{if } V < \varphi + \lambda, \ R = 1 
\end{cases} $$

Throughout the paper coefficients of internal variables in structural equations are indicated by $\alpha$, while coefficients of external variables are indicated by $\beta$. Here $\alpha_Y > 0$ implies that the agent will have a better outcome if the treatment is administered than if it is not. Further, $\beta_Y < 0$ implies that type-0 agents are likely to have better outcomes than type-1 agents whether or not they are treated. Finally, $\lambda > 0$ implies that type-0 agents are likelier to get treatment than type-1 agents.

We now ask in what sense, if any, it is possible to identify the effect of treatment on outcomes in this model. Without qualification of the question it is not possible to do so. An intervention on the treatment variable—replacing $T =$
0 with $T = 1$—can result from an intervention in either $R$ or $V$, and the effect of the intervention on $Y$ depends on which is the case, even though $\Delta T = 1$ in both cases. (To see that the effects on $Y$ are different it is sufficient to note that if the intervention in $T$ is caused by $V$ the effect on $Y$ does not depend on $\beta_{Y R}$, whereas if the intervention is on $R$ then $\beta_{Y R}$ does enter the expression for the effect on $Y$). Thus specifying $\Delta T = 1$ does not provide enough information about the intervention to determine $\Delta Y$. In particular, we cannot characterize $\alpha_{Y T}$, or any other parameter, as parametrizing the effect of $\Delta T$ on $\Delta Y$. In this situation causation is not implementation neutral. Causation is implementation neutral when all interventions on external variables that lead to a given change in the cause variable induce the same change in the effect variable, implying that one can treat the cause variable as if it were external regardless of what the underlying intervention is.

As a semantic point one could make a case for restricting the use of statements like “the causal effect of $\Delta T$ on $\Delta Y$ is ...” to settings where the conditions for implementation neutrality are satisfied, since the magnitude of causal effects can be associated with a parameter (or variable in the case of nonlinear models) only in that case. Doing so, however, would constitute a radical departure from existing usage, under which the relation between two variables is causal if all external variables that affect the cause variable also induce a change in the effect variable. This is a weaker condition. To ensure a clear distinction between the two concepts of causation we will use the term “IN-causation” to distinguish implementation-neutral causation from causation that is not necessarily implementation neutral. Thus when the effect is causal but not IN-causal no parameter (or variable) measuring the strength of causation is defined. When it is IN-causal there exists a parameter or variable that measures the strength of causation.

The special case of $\beta_{Y R} = \lambda = 0$ in the model just set out yields a model in which treatments are assigned according to a random process. Under random treatments the implementation neutrality condition is satisfied. This is so because by assumption the random variable that generates the assignment does not directly affect any variables other than the treatment variables. However, if treatment is not determined randomly one has to consider the possibility that external variables that affect the treatment also directly affect the outcome variable, as in the example just presented under general parameter values. As Heckman has pointed out, although in a different context, randomization is sufficient for unbiased estimation of causal parameters, but it is not necessary. Failure of randomization does not necessarily imply failure of implementation neutrality.

### 2 The Treatment Evaluation Approach

Under the treatment evaluation analytical framework there is no analogue to eq. (2) representing the generation of the treatment variable. Instead, the practice is directly to specify two outcome variables $Y(1)$ and $Y(0)$, representing the
outcomes if the treatment is and is not applied. Much is made of the fact that either $Y(1)$ or $Y(0)$ for an individual agent is necessarily a counterfactual, and therefore cannot be directly observed (Rubin [1974], [1978]). If $T$ IN-causes $Y$ there is no problem with defining $Y(1)$ and $Y(0)$ in this way, since in that case $Y(T)$ is unambiguously defined for both values of $T$ by the definition of implementation neutrality.\(^1\) In the contrary case, however, the values $Y(1)$ and $Y(0)$ are not implied by the hypothesized intervention, implying that the effect of $T$ on $Y$ cannot be identified with $Y(1) - Y(0)$. In our setting $Y(1)$ and $Y(0)$ depend on $R$ and $V$, but under the treatment evaluation approach these variables have no analogue, so the validity of the analysis is restricted to the case in which causation is implementation neutral. This rules out most of the cases of interest.

3 Other Statements of Causation

We have seen that due to failure of implementation neutrality one cannot quantify the effect of $T$ on $Y$ in the model just presented, at least without imposing restrictions. Other causal statements do not require implementation neutrality, implying that they are valid in the example as written, rather than only in special cases. Most simply, we have that the effect of $T$ on $Y$ conditional on $R$ is well defined without restrictions on model coefficients, and equals $\alpha_{YT}$ regardless of whether $R = 0$ or $R = 1$.

Analysts sometimes consider the effect of treatment on the treated, or on the untreated. Bayes’ rule allows an easy calculation of the proportions of $R = 0$ and $R = 1$ agents in the treated and untreated populations; these differ because the $R = 0$ agents are more likely to be treated than the $R = 1$ agents. Nevertheless, the effect of the treatment on the treated or the untreated is not well defined, again due to the failure of implementation neutrality. Just as in the case with no conditioning, restricting the population to the treated does not alter the fact that going from $T = 0$ to $T = 1$ could occur as a result of an intervention on either $R$ or $V$, and these have different implications for $Y$. Again, the simplest way to see this is to consider an agent with $R = 0$ and $V$ satisfying $\varphi \leq V < \varphi + \lambda$. From eq. (2), this agent would be treated. Now consider an intervention that results in the same agent being untreated. This will occur if $V$ is reduced to a level below $\varphi$, or if $R$ is changed to 1. The effect on $Y$ of the former intervention does not depend on $\beta_{YR}$, while the latter intervention does depend on $\beta_{YR}$. Clearly characterizing an intervention as a change of $T$ from 1 to 0 is not sufficient to determine the consequence for $Y$.

\(^1\)Some papers appear to adopt a hybrid combination of the econometric approach and the treatment evaluation approach. For example, Imbens and Angrist [1994] explicitly included an equation relating $T$ to its external determinants, but also assumed that $Y(0)$ and $Y(1)$ are well defined, as is the case only if $T$ IN-causes $Y$. 

4
4 Implementation Neutrality and Instrumental Variables

Many evaluations of treatment effects have to consider the possibility of correlation between the treatment variable and an unobserved error. Existence of this correlation creates a presumption that ordinary least squares estimates of treatment effectiveness will be biased and inconsistent. The standard treatment is to use an instrumental variables estimator rather than ordinary least squares. If the instrument is correlated with the treatment but not with the error the problem of bias is eliminated. In this section we consider the role of instrumental variables estimators in empirical estimation of causal parameters.

Angrist’s [1990] paper evaluating the effects of military service on lifetime earnings is an outstanding example of such use of instrumental variables. The problem is that ordinary least squares estimates of the effect of military service on lifetime earnings appear to be biased because veteran status is correlated with such variables as ability to earn a high income in civilian employment, which in turn is a correlate of lifetime earnings. Angrist’s solution was to use measures of eligibility for conscription implied by the Vietnam era lotteries as an instrument. Whether or not an agent is eligible for conscription is correlated with whether or not he served in the military—the treatment—but, arguably, not with other determinants of lifetime earnings. This justification for draft eligibility as an instrument in estimating the parameter Angrist associated with the effect of veteran status on earnings is entirely persuasive, although Angrist did list some possible remaining sources of bias.

A simplified version of Angrist’s model is

\[ E = \beta_E V + U, \]  \hspace{1cm} (3)

where \( E \) is lifetime earnings, \( V \) is a 0-1 variable measuring whether an agent served in the military, and \( U \) is an error. Here \( \beta_E \) is a parameter representing the causal effect of \( V \) on \( E \), reflecting the assumption that \( V \) can be taken as external. The assumed correlation between \( V \) and \( U \) motivates use of the instrumental variables estimator: “The justification for estimation of the effects of military service in this manner is clear: it is assumed that nothing other than differences in the probability of being a veteran is responsible for differences in earnings by draft-eligibility status.” (p. 320).

The validity of this justification for Angrist’s estimation of the effects of military service is in fact far from clear. The problem is not with the use of instrumental variables, which as asserted does produce an unbiased estimate of the parameter under consideration, given that the model is not misspecified. The problem is with interpreting \( V \) as an external variable despite the existence of an uninterpreted correlation between \( V \) and \( U \). Determining whether \( \beta_E \) can be taken to be a causal parameter, as asserted, requires deconstructing the correlation between \( V \) and \( U \). To do this we introduce a new variable into the model. Let \( A \) represent an unobservable measure of an agent’s ability to earn a high income in civilian employment, which is correlated with both \( E \) and \( V \).
The problem is how to specify which of $A$ and $V$ is external. For our purpose there are two possibilities. First, consider what Angrist characterized as the simplest interpretation, that agents in military service accumulate human capital at a different rate from those in civilian employment, resulting in different future incomes when they compete in civilian job markets against nonveterans. Under this interpretation the augmented model is written

$$E = \beta_{EV}V + \alpha_{EA}A + U$$

(4)

$$A = \beta_{AV}V + W.$$  

(5)

The external variables are $U$, $V$ and $W$, and they are assumed unobserved and uncorrelated.

Specifying $V$ as an external variable implies that there is no problem with interpreting the correlation between $V$ and $E$ as reflecting the causal effect of $V$ on $E$. The causal coefficient for the effect of $V$ on $E$ is $\Delta E/\Delta V = \beta_{EV} + \alpha_{EA}\beta_{AV}$, which is estimated consistently by ordinary least squares. The question is how to interpret $\beta_{EV}$, the parameter Angrist sought to estimate. For generic parameter values $\beta_{EV}$ does not capture the causal effect of $V$ on $E$, inasmuch as its sample counterpart can be interpreted as providing a biased estimate of $\Delta E/\Delta V$.

However, $\beta_{EV}$ can be interpreted as representing the causal effect of $V$ on $E$ holding constant $A$. As such, the intervention associated with $\beta_{EV}$ consists of an intervention consisting of a hypothetical change in $V$ accompanied by a simultaneous offsetting intervention on $W$ of magnitude just sufficient to cancel out the effect of $V$ on $A$. It is far from clear how to interpret such a complicated intervention intuitively (it is argued in LeRoy [2015] that interventions that involve conditioning on an internal variable presume a functional relation among external variables that contradicts their assumed specification as external, and therefore should be disallowed). Thus the conclusion is that with veteran status specified as an external variable the causal relation between veteran status and lifetimes earnings should be identified with unconditional causation. The correct estimator of the causal effect of veteran status on earnings is the ordinary least squares regression coefficient, not the instrumental variables estimator.

An alternative setup that produces a model in which $V$ and $U$ are correlated in eq. (3) is that $A$, not $V$, is the external variable, so that variations in civilian earnings ability induce changes in veteran status and also are correlated with subsequent earnings levels. Thus the model becomes

$$E = \alpha_{EV}V + \beta_{EA}A + U$$

(6)

$$V = \beta_{VA}A + W,$$  

(7)

with $U$, $A$ and $W$ assumed uncorrelated. With $V$ now an internal variable we have to ask whether $V$ IN-causes $E$, so as to be able to consider the effect of $V$ on $E$. It does not: the two external variables $A$ and $W$ that determine
$V$ have different effects on $E$ holding constant $\Delta V$. Thus no parameter can be interpreted as reflecting the unconditional effect of $V$ on $E$.

A different conclusion results if we consider the effect of $V$ on $E$ conditional on $A$, rather than the unconditional effect of $V$ on $E$ as above. A case could be made that the conditional effect is in fact of primary interest: an individual considering whether to enlist in the military would take the view that his ability to earn an income in civilian employment is whatever it is, indicating the appropriateness of conditioning on $A$. With the conditional effect there is no violation of implementation neutrality: $V$, being essentially equivalent to $W$, IN-causes $E$.\(^2\) The coefficient $\alpha_{EV}$ gives the causal coefficient of $V$ on $E$ and, as Angrist showed, this coefficient is estimated without bias using draft status as an instrumental variable.

5 Conclusion

The examples underline the importance of specifying explicitly how treatments are generated in the data used to appraise treatment effectiveness, rather than working directly with uninterpreted correlations. If randomization is available there are no problems. Even in the absence of randomization there may be ways to achieve estimates that are unbiased, but whether this is so depends on what assumptions the analyst is willing to specify for his model.

We have seen that the conditions required for implementation neutrality depend on the causal statement that is envisioned: some statements of causation are invalidated due to failure of implementation neutrality, while others carry over. In our examples we have provided examples of each. Analysts need to distinguish among alternative possible causal statements and avoid those that are invalid in the models they specify.

References


\(^2\)Because $A$ is specified to be external, conditioning on $A$ does not involve a violation of the assumed variation-free character of the external variables. In contrast, we noted above that conditioning on internal variables necessitates constraints on the assumed intervention on external variables, leading to difficulties of interpretation.


