Economics and Philosophy, **34** (2018) 45–52 © Cambridge University Press
doi:10.1017/S0266267117000219 First published online 26 September 2017
cambridge.org/eap

IMPLEMENTATION NEUTRALITY AND TREATMENT EVALUATION

STEPHEN F. LEROY*

Abstract: Statisticians have proposed formal techniques for evaluation of treatments, often in the context of models that do not explicitly specify how treatments are generated. Under such procedures they run the risk of attributing causation in settings where the implementation neutrality condition required for causal interpretation of parameter estimates is not satisfied. When treatment assignments are explicitly modelled, as economists recommend, these issues can be formally analysed, and the existence (or lack thereof) of implementation neutrality, and therefore quantifiable causation, can be determined. Examples are given.

Key Words: Causation, Treatment evaluation, Implementation neutrality, Identification

Statisticians associated with a number of fields – 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 biased estimators of causal effects. 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.

* Department of Economics, University of California, Santa Barbara 93106, CA, USA. Email: leroy@ucsb.edu

As many have noted, noneconomists resist this approach. They instead propose mechanical algorithms that purportedly make possible diagnosis of causal relations, in particular of treatment evaluation, without committing to any particular representation of the assignment mechanism (Spirtes *et al.*, 1993; Pearl 2001). 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 show 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 unjustified except in special cases in which conditions for implementation neutrality (defined below) are satisfied.[1] For other causal statements, however, the conditions for implementation neutrality are satisfied in the model as specified, so there is no problem with causal interpretations. Distinguishing between these two cases enables the analyst to determine which causal statements are valid in a given model and which are not.

The analysis is then recast in the framework used in the treatment evaluation literature (Section 2). We show that this alternative representation implicitly assumes satisfaction of an implementation neutrality property the conditions for which are not explicitly modelled, resulting in the apparent validity of causal statements that may be unjustified.

A second example, discussed in Section 3, involves an evaluation of instrumental variables estimators in settings where the assumptions underlying ordinary least squares estimators are violated.

1. THE ECONOMETRIC APPROACH

An econometric model intended to generate causal statements requires explicit specification of variables and a labelling of each variable as external or internal, so as to make clear which variables the model is intended to explain. When there exists a unique equilibrium the model implies the existence of a vector function mapping external variables to equilibrium values of internal variables. Causal orderings can be determined by analysing this function.

In our first example the external variables consist of three random variables: R, u and v. Throughout this paper variables denoted by a capital letter are assumed to be observable, and those denoted by a lower-case letter are unobservable. R is a binary random variable that

[1] The basic ideas involving implementation neutrality were presented elsewhere (LeRoy 2016).

https://doi.org/10.1017/S0266267117000219 Published online by Cambridge University Press

is interpreted as an agent's race. R is assumed to take on values 0 or 1 with given probabilities. The errors u and v are unobservable real-valued random variables with given distributions. The external variables are independently distributed, consistently with the assumption that they are not linked by equations of the model (otherwise they would not be external). 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 model consists of the following equations:

$$Y = \alpha_{YT} T + \beta_{YR} R + u \tag{1}$$

$$T = \begin{cases} 1 & \text{if } \beta_{TR} R + v \geq 0 \\ 0 & \text{if } \beta_{TR} R + v < 0. \end{cases} \tag{2}$$

Throughout the paper coefficients of internal variables are indicated by α, while coefficients of external variables are indicated by β. Here $\beta_{TR} > 0$ implies that type-1 agents are likelier to get treatment than type-0 agents, and this plus $\alpha_{YT} > 0$ and $\beta_{YR} > 0$ imply that type-1 agents are likely to have better outcomes than type-0 agents.

We now ask in what sense, if any, is it possible to evaluate quantitatively the effect of treatment on outcomes in this model. It is not possible to do so. A hypothetical alteration of the treatment variable – replacing $T = 0$ with $T = 1$ – can result from an intervention on 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.

The simplest way to see this is to consider an agent with $R = 0$ and $-\beta_{TR} < v < 0$. From eq. (2) this agent would not be treated. Now consider an intervention that results in an agent with the same u being treated. This could occur either because v is increased to a level greater than 0, or because R is changed to 1. The effect on Y is α_{YT} in the first case, or $\alpha_{YT} + \beta_{YR}$ in the second. Clearly, characterizing an intervention as a change of T from 0 to 1 is not sufficient to determine the consequence for Y: specifying $\Delta T = 1$ does not provide enough information about the intervention to determine ΔY. In particular, we cannot characterize α_{YT}, or any other parameter, as parameterizing the effect of T on Y.

In this situation causation is not *implementation neutral*: different interventions that implement the same ΔT may lead to different values of ΔY, so that the effect of T on Y is not well defined. Causation is implementation neutral when all interventions that implement a given ΔT have the same effect on ΔY.

As a semantic point one could make a case for restricting the use of statements like 'the causal effect of ΔT on ΔY is ...' to settings where the conditions for implementation neutrality are satisfied, since

https://doi.org/10.1017/S0266267117000219 Published online by Cambridge University Press

the magnitude of causal effects can be associated with a parameter of the model (or variable in the case of nonaffine 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, a weaker condition. To ensure a clear distinction between the two concepts we will use the term 'causation' with its usual meaning, and will reserve the term 'IN-causation' for the case in which causation is implementation neutral. When the relation is IN-causal and the model is affine there exists a parameter that measures the strength of causation. When the relation is causal but not IN-causal no parameter measuring the strength of causation is defined.

2. THE TREATMENT EVALUATION APPROACH

Under the treatment evaluation analytical framework the practice is directly to specify two outcomes $Y(1)$ and $Y(0)$, representing the outcomes for a particular agent if the treatment is or is not applied. Much is made of the obvious 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, in a model that specifies T to be internal, 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. In the contrary case, however, the values $Y(1)$ and $Y(0)$ are not uniquely characterized by the hypothesized intervention on T , implying that the effect of T on Y cannot be identified with $Y(1) - Y(0)$, as would otherwise be possible (in affine models). In our example $Y(1)$ and $Y(0)$ depend on R and v, but under the treatment evaluation approach these latter variables do not appear. Accordingly, the validity of the analysis is restricted to the case in which causation is implementation neutral despite the fact that nothing in the model implies that this condition is satisfied.

3. 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 are inconsistent. The standard procedure is to use an instrumental variables estimator rather than ordinary least squares. If the instrument is correlated with the treatment variable but not with the error the problem of inconsistency is eliminated. In this section we specialize the discussion to consider the role of instrumental variables estimators in empirical estimation of causal parameters in the presence of correlation among explanatory variables.

https://doi.org/10.1017/S0266267117000219 Published online by Cambridge University Press

In general the parameters associated with IN-causation are not identified without an assumption that unobservable external variables are uncorrelated with each other and with observed external variables. The questions are whether IN-causality is preserved under instrumental variables estimation and whether the identification problem persists. Resolving these questions involves incorporating the instrument in the model in such a way that all variables characterized as external are uncorrelated. Then one ascertains the causal ordering in the modified model, the questions being whether causation is implementation-neutral and, if so, whether the associated parameter is identified.

Reformulating models so as to resolve the correlations that complicate empirical estimation of causal parameters requires that the model-builder take a stand on why the variables are correlated: if in the model (1) described above T and u are correlated the model-builder must introduce a constant λ and a new variable z and write $T = \lambda z + u$ (this, of course, involves respecifying u). He must then specify which two of T, z and u are external, and therefore may be taken to be uncorrelated. If z and u are defined to be external, T becomes an internal variable, while if z and T are specified to be external, then u is internal. So reformulated the model inherits the properties of models with uncorrelated external variables. In general the IN-causal ordering depends on how correlations among external variables are resolved.

Angrist's (1990) paper evaluating the effects of military service on lifetime earnings provides a setting in which these difficulties can be explored. One can estimate the effect of military service on the lifetime earnings Y of veterans and non-veterans by running the regression

$$Y = \beta_{YV} V + u, \tag{3}$$

where V is a dummy for military service. If V is external there is no problem with asserting that V IN-causes Y, with β_{YV} measuring the magnitude of the effect. The problem is that an ordinary least squares estimate of β_{YV} is biased to the extent that veteran status is correlated with such unobserved variables as ability to earn a high income in civilian employment, which in turn is an explanatory variable for lifetime earnings. Thus V and u are correlated, so the population parameter β_{YV} is not identified.

Angrist's solution was to use a measure E of eligibility for conscription as an instrument in estimating β_{YV}. E was specified to consist of the number associated with each agent under the draft lottery in the Vietnam war. Whether or not an agent is likely to be drafted based on his lottery number is correlated with whether or not he served in the military – the treatment – but, arguably, not with other determinants of lifetime earnings. This, it is suggested, establishes the suitability of E as an instrument.

https://doi.org/10.1017/S0266267117000219 Published online by Cambridge University Press

This justification for draft eligibility as an instrument in estimating the parameter Angrist associated with the effect of veteran status on earnings seems persuasive, but the informal treatment of the correlation between V and u is problematic. Investigating this difficulty involves dispensing with the purely verbal treatment of draft eligibility and earnings ability in favour of working with a model that incorporates these variables explicitly.

Let (unobservable) a represent an agent's ability to earn a high income in civilian employment. The new variables E and a are not part of the original formal model, consisting of eq. (3). We now expand that model to incorporate them, and use the expanded model to deconstruct the correlation between V and u. The problem is to specify which variables are external in the expanded model. For the purpose of the present discussion there are two possibilities. First, consider what Angrist characterized as the simplest specification for why military service affects lifetime earnings: 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 can be written

$$Y = \alpha_{YV} V + \alpha_{Ya} a + u \tag{4}$$

$$a = \alpha_{aV} V + w \tag{5}$$

$$V = \begin{cases} 1 & \text{if } \beta_{VE} E + z \geq 0 \\ 0 & \text{if } \beta_{VE} E + z < 0 \,. \end{cases} \tag{6}$$

The external variables here are E, u, w and z. These are assumed to be distributed independently. Eq. (5) expresses the dependence of earnings ability on veteran status, while eq. (6) specifies that veteran status depends on eligibility for the draft. The model can be parameterized so that the implied joint distribution of Y and V is the same as in the original model (3).

Here, as in the original regression (3), α_{YV} cannot be estimated consistently by an ordinary least squares regression of Y on V because V is correlated with a, which is a component of the error. Instrumental variables also produces an inconsistent estimate of α_{YV} because E is correlated with a, using eqs (5) and (6). Finally, we have that V does not IN-cause Y, implying that α_{YV} does not represent an IN-causal influence. The appeal to instrumental variables to produce a consistent estimator of an IN-causal parameter fails.

Instead of having veteran status IN-causally prior to earnings ability, we could reverse the causation and specify that earnings ability IN-causes veteran status, so that agents are more or less likely to join the armed

https://doi.org/10.1017/S0266267117000219 Published online by Cambridge University Press

forces according to their earnings ability in civilian employment. A model that reflects this respecification is the following:

$$Y = \alpha_{YV} V + \alpha_{Ya} a + u \tag{7}$$

$$V = \begin{cases} 1 \text{ if } \beta_{VE} E + \beta_{Va} a + z \geq 0 \\ 0 \text{ if } \beta_{VE} E + \beta_{Va} a + z < 0, \end{cases} \tag{8}$$

where a, E, u and z are external, and assumed uncorrelated. In this setting an ordinary least squares regression of Y on V yields an inconsistent estimate of α_{YV} for the same reason as above. However, using E as an instrument in an instrumental variables estimate yields a consistent estimator of α_{YV}: E, being external, is uncorrelated with the error, but is correlated with V. Thus E is a valid instrument for α_{YV}, contrary to the earlier case. However, it remains true that V does not IN-cause Y, again due to the presence of a, an element of the external set for V, as a separate explanatory variable for Y in eq. (7). Thus α_{YV}, although now consistently estimated, does not represent the IN-causal effect of V on Y. Again the instrumental variables estimator, although consistent, does not yield an estimator of IN-causation.

We see that recasting the model so as to eliminate uninterpreted correlations may not preserve IN-causal orderings. If not, causal parameters are not well defined, so there is nothing to estimate. Either way, it would seem, the potential role of instrumental variables estimators is unclear.

4. CONCLUSION

The examples underline the importance of specifying explicitly how treatments are generated in the data used to appraise treatment effectiveness, rather than attempting to work directly with uninterpreted correlations. 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 instances of each. Analysts need to distinguish among alternative possible causal statements and avoid those that are invalid in the models they specify.

ACKNOWLEDGEMENTS

I have received helpful comments from Hrishikesh Singhania, Richard Startz and Douglas Steigerwald.

REFERENCES

Angrist, J. D. 1990. Lifetime earnings and the Vietnam era draft lottery. *American Economic Review* 80: 313–336.

Heckman, J. J. 2001. *Econometrics Counterfactuals and Causal Models*. Chicago, IL: University of Chicago.
LeRoy, S. F. 2016. Implementation neutrality and causation. *Economics and Philosophy* 32: 121–142.
Pearl, J. 2001. *Causality*. Cambridge: Cambridge University Press.
Rubin, D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66: 688–701.
Rubin, D. 1978. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 7: 34–58.
Spirtes, P., C. Glymour and R. Schienes. 1993. *Causation, Prediction and Search*. Cambridge, MA: MIT Press.

BIOGRAPHICAL INFORMATION

Stephen F. LeRoy is Professor of Economics Emeritus at the University of California, Santa Barbara. He is coauthor, with Jan Werner, of *Principles of Financial Economics*. He has been a research economist in the Federal Reserve System, and has served on the faculties of the University of Chicago, California Institute of Technology, University of California, Berkeley and other universities. His research focuses on economic theory and finance.