Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Comment

James J. Heckman

To cite this article: James J. Heckman (1996) Comment, Journal of the American Statistical Association, 91:434, 459-462, DOI: 10.1080/01621459.1996.10476904

To link to this article: https://doi.org/10.1080/01621459.1996.10476904

Comment

James J. HECKMAN

Angrist, Imbens, and Rubin (AIR) apply the method of instrumental variables (IV) to estimate the local average treatment effect (LATE) of Imbens and Angrist (1994). Application of IV is routine, even for evaluation models with heterogeneous responses to treatment, so there is nothing novel or controversial about the method.

LATE is a controversial parameter because it is defined for an unobservable subpopulation. Its use as an evaluation parameter thus is of questionable value. More controversial yet is AIR's mischaracterization of the current state of knowledge about econometric models of simultaneity and selectivity. Econometrics has moved well beyond the (1943) model of Haavelmo and the simpler cases of the dummy endogenous variable model of Heckman (1978) to which the authors confine their attention in comparing their approach to econometric methods. It is interesting to contrast their commentary on econometric models for simultaneous equations with the commentary of scholars of causal analysis in science. For example, the distinguished philosopher Nancy Cartwright (1989) demonstrated how econometric methods for simultaneous equations elucidate causality, provide a coherent scheme for generating counterfactuals, and shed light on controversies in quantum mechanics.

This comment makes four points:

1. The "Rubin model" is a version of the widely used econometric switching regression model (Maddalla 1983; Quandt 1958, 1972, 1988). The Rubin model shares many features in common with the Roy model (Heckman and Honoré 1990; Roy 1951) and the model of competing risks (Cox 1962). It is a tribute to the value of the framework that it has been independently invented by different disciplines and subfields within statistics at different times.

2. Contrary to remarks by AIR, econometric work on simultaneous equations allows for variable responses to treatment, does not rely on arbitrary distributional assumptions, develops IV estimation methods for these models, examines the assumptions required to justify IV, and demonstrates that the assumptions required to use IV in the general case are very strong. This analysis is conducted within the context of clearly specified models of outcomes and regime selection that are motivated by behavioral theory. Econometricians make *weaker* mean independence assumptions rather than the strong independence assumptions made by AIR to identify their parameter.

3. The independence assumptions invoked by AIR are based on unspecified and implicit behavioral assumptions. These assumptions about behavior are very unattractive once they are clearly stated.

4. Econometric policy evaluation is designed to produce many counterfactuals from a common set of behavioral functions. Conditions required to nonparametrically identify this common set of functions are presented in the econometrics literature.

1. SWITCHING REGRESSION MODELS AND THE "RUBIN MODEL"

Counterfactuals are at the heart of any scientific study. Galileo was perhaps the first to use the thought experiment and the idealized method of controlled variation to define causal effects. Economists have used this method from the time of Alfred Marshall, who repeatedly used the principle of controlled variation in his *ceteris paribus* clauses. Since Haavelmo (1944), modern econometrics has been devoted to the construction and estimation of a broad array of counterfactuals. The construction of counterfactual states is the essence of econometric policy analysis (see Lucas and Sargent 1981).

Economists have used the following specific model of potential outcomes for at least 25 years. It has its origin in classical models of choice among discrete outcomes in mathematical psychology pioneered by Thurstone in the 1920s. (See Falmagne 1985 for an extensive bibliography; see also McFadden 1974 or Quandt 1972). There are two possible regimes, "0" or "1." (The generalization to an arbitrary number is trivial. I use two regimes to conform to the setup of AIR.) Associated with each regime is an outcome (Y^0 or Y^1). There is a rule that selects regimes. Let $D = 1$ if "1" is selected; $D = 0$ otherwise. Variables $\mathbf{X}$ determine outcomes in the following sense: $E(Y^1|\mathbf{X}) = \mu^1(\mathbf{X}), E(Y^0|\mathbf{X}) = \mu^0(\mathbf{X})$. Variables $\mathbf{X}$ and $\mathbf{Z}$ determine D in the following sense: $\Pr(D = 1|\mathbf{X}, \mathbf{Z})$ is a nontrivial function of both $\mathbf{X}$ and $\mathbf{Z}$. The "mysterious errors" that AIR censure in econometric work (see the last paragraph of their sec. 2) are (U^0, U^1) defined as $U^0 = Y^0 - E(Y^0|\mathbf{X}), U^1 = Y^1 - E(Y^1|\mathbf{X})$. (There are more general models, as considered in Heckman and Robb 1985, but for the sake of brevity I consider only the simplest case.) These "errors" are well defined as long as $E(Y^0) < \infty, E(Y^1) < \infty$. If (U^0, U^1) are independent across observations "SUTVA" holds.

Only one regime is observed at any time. Let Y be the observed outcome. $\Delta = Y^1 - Y^0$. Then

$$Y = DY^1 + (1 - D)Y^0 = Y^0 + D(Y^1 - Y^0) = Y^0 + D\Delta, \tag{1}$$

James J. Heckman is the Henry Schultz Distinguished Service Professor, Department of Economics, University of Chicago. This research was supported by National Science Foundation Grant SBR-93-048-0211 and a grant from the Russell Sage Foundation.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Applications and Case Studies

so

$$Y = \mu^0(\mathbf{X}) + D(\mu^1(\mathbf{X}) - \mu^0(\mathbf{X}) + U^1 - U^0) + U^0. \quad (2)$$

The "gain" from going from "0" to "1" is $\Delta = Y^1 - Y^0$. This is what AIR and previous authors in econometrics call the "causal effect" of D on Y. If superscript "1" refers to demand and "0" to supply, we obtain the disequilibrium markets model of Quandt (1972, 1988) where

$$D = 1(Y^1 \leq Y^0); = 0 \quad \text{otherwise.}$$

"1" is the logical indicator variable ($1(a) = 1$ if a holds). Letting Y^1 be the market wage and Y^0 the value of nonmarket time, we obtain the model of the value of wages and of nonmarket time presented by Gronau (1974) and Heckman (1974), where $D = 1(Y^1 \geq Y^0)$. The same model appears in studies of search unemployment (Flinn and Heckman 1982). If Y^1 and Y^0 are times of death in a competing risk setup, then $D = 1(Y^1 \leq Y^0)$. Amemiya (1985), Heckman and Robb (1985), and Willis and Rosen (1979) considered more general models for D. Applications of this basic model of potential outcomes are legion in economics. Analyses are available both for the case where D is observed and where it is not. (See the numerous references in Amemiya 1985.)

2. MEAN EFFECT OF TREATMENT ON THE TREATED: A "CAUSAL" PARAMETER

A basic program evaluation parameter widely used in economics is the mean effect of treatment on the treated:

$$E(Y^1 - Y^0|D = 1, \mathbf{X}) = E(\Delta|D = 1, \mathbf{X}). \quad (3)$$

This parameter tells us what, on average, persons of characteristics $\mathbf{X}$ who actually participate in regime 1 gain from a switch from regime "0" to "1." Unlike LATE, this parameter is defined for an observable subpopulation. It is a parameter that corresponds more closely to the coefficient on the dummy variable in nonlinear simultaneous equation (2) with variable treatment effect than does LATE. Using means of nonparticipants to estimate $E(Y^0|D = 1, X)$ gives rise to selection bias if $E(Y^0|D = 1, \mathbf{X}) - E(Y^0|D = 0, \mathbf{X}) \neq 0$. Equivalently, $E(U^0|\mathbf{X}, D = 1) \neq E(U^0|\mathbf{X}, D = 0)$. I discuss identification and estimation of this parameter because there is a vast literature on it in economics, and the lessons from an analysis of this parameter apply directly to LATE.

Equation (2) is a *variable effect* or random coefficient model. The response to treatment (the term multiplying D) varies among persons with identical $\mathbf{X}$, unless $U^1 = U^0$. That special case is the dummy endogenous variable model of Heckman (1978) discussed by AIR. We may reparameterize Equation (2) in terms of $E(\Delta|D = 1, \mathbf{X})$. The reparameterization writes

$$Y = \mu^0(\mathbf{X}) + D(E(\Delta|\mathbf{X}, D = 1)$$
$$+ \{U^0 + D(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1))\}.$$

The term in braces is an interpretable "error term." If D were orthogonal to this term, then simple means for each $\mathbf{X}$ group would identify the parameter of interest. (For each $\mathbf{X}$, subtract the mean for $D = 0$ from the mean for $D = 1$). Selection bias makes D nonorthogonal to U^0. D is orthogonal to the second error component in the braces ($E(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1)|\mathbf{X}, D = 1) = 0)$), but not to the first component.

The method of IV has been applied to identify this parameter under general conditions (see Heckman and Robb 1985, 1986). Suppose that Z is distinct from X (i.e., does not appear directly in (2)) and satisfies the following *mean independence* conditions:

$$E(U^0|\mathbf{X}, \mathbf{Z}) = 0 \quad \text{(A-1)}$$

$$E(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1)|\mathbf{X}, \mathbf{Z}, D = 1) = 0. \quad \text{(A-2)}$$

An alternative and equivalent way to write condition (A-2) is:

$$E(\Delta|\mathbf{X}, \mathbf{Z}, D = 1) = E(\Delta|\mathbf{X}, D = 1). \quad \text{(A-2')}$$

Finally, restate the condition that both $\mathbf{Z}$ and $\mathbf{X}$ determine D as an assumption:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}) \neq \Pr(D = 1|\mathbf{X}) \quad \text{and} \quad \Pr(D = 1|\mathbf{X}, \mathbf{Z}) \quad \text{(A-3)}$$

is a nontrivial function of $\mathbf{Z}$. Then

$$E(Y|\mathbf{X},\mathbf{Z}) = \mu^0(\mathbf{X}) + E(\Delta|\mathbf{X},D = 1)\Pr(D = 1|\mathbf{X},\mathbf{Z}). \quad (4)$$

If for each $\mathbf{X}$ there are at least two distinct values of $\mathbf{Z}$ such that $\Pr(D = 1|\mathbf{X}, \mathbf{Z}') \neq \Pr(D = 1|\mathbf{X}, \mathbf{Z}'')$, then we may evaluate (4) at all values of X to obtain

$$E(\Delta|\mathbf{X}, D = 1) = \left(\frac{E(Y|\mathbf{X}, \mathbf{Z}') - E(Y|\mathbf{X}, \mathbf{Z}'')}{\Pr(D = 1, \mathbf{X}, \mathbf{Z}') - \Pr(D = 1|\mathbf{X}, \mathbf{Z}'')} \right). \quad (5)$$

If for some $\mathbf{X}$ values there are not distinct values for $\Pr(D = 1|\mathbf{X}, \mathbf{Z})$ for two or more values of Z, then the parameter is not identified at those values. Replacing population objects with sample mean analogs produces the IV estimator.

Observe that only mean independence is required—not full independence, as assumed by AIR. AIR are able to test their identifying assumptions because they invoke much stronger conditions than are required to identify their parameter. Minimal identifying assumptions cannot be tested. (Heckman and Robb 1985). Note further that no arbitrary and untestable monotonicity condition is needed—just a condition that guarantees that the denominator of (5) is not zero for the particular value of $\mathbf{X}$. Parenthetically, monotonicity is not required in classical discrete choice theory either. Also, even in the original dummy endogenous variable analysis it is recognized that the second assumption of AIR's Equation (4) is not needed to apply IV.

3. ARE THE ASSUMPTIONS VALID?

A central focus in modern econometrics is the development of explicit behavioral models relating the "errors" and choices made by agents. This is critical to developing and justifying any econometric evaluation strategy. Therefore, it is surprising to read in AIR that econometricians do not clearly state assumptions like (A-1)–(A-3) or worry about the justification for them. Hansen and Sargent (1991), and Heckman and Robb (1985, 1986) are just some of the authors who have built explicit behavioral models to justify exogeneity and noncausality assumptions.

Assumption (A-1) is conventional; Assumption (A-2) (or A-2′) is not. (A-2) is satisfied if $U^1 - U^0 = 0, (Y^1 - Y^0 = \mu^1(\mathbf{X}) - \mu^0(\mathbf{X}))$, so that conditional on $\mathbf{X}$ there is no treatment response heterogeneity. It is also satisfied in the case of heterogeneous response to treatment when the rule governing participation in a regime conditional on $\mathbf{Z}$, and $\mathbf{X}$ does not depend on $Y^1 - Y^0$:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}, Y^1 - Y^0) = \Pr(D = 1|\mathbf{X}, \mathbf{Z}), \quad \text{(6a)}$$

provided that $Y^1 - Y^0$ conditional on X is not perfectly forecastable by $\mathbf{Z}$. This is a Granger noncausality condition routinely used in econometrics and explicitly presented in this context by Heckman and Robb (1985, 1986). Alternatively, in terms of the "mysterious" unobservables to which AIR object, the condition is:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}, U^1 - U^0) = \Pr(D = 1|\mathbf{X}, \mathbf{Z}), \quad \text{(6b)}$$

provided that $U^1 - U^0$ conditional on $\mathbf{X}$ is not perfectly forecastable by $\mathbf{Z}$. AIR call this non-causality condition "ignorability." If $U^1 - U^0$ is perfectly forecastable by $\mathbf{Z}$, conditional on $\mathbf{X}$, then (A-2) would be violated.

In general, the extra conditioning on $\mathbf{Z}$ causes (A-2) to be violated although it is trivially satisfied only if conditioning is done on X and D. The behavioral assumption justifying (6a) and (6b) requires that the relevant decision makers do not make decisions about which regime is selected using information on the outcomes of the regime that cannot be forecast by X and Z. In most situations, persons making decisions have more information about the outcomes than the statisticians studying them. This makes assumption (6a) or (6b) questionable in such cases.

This assumption is definitely not satisfied in the competing risks model, in the Gronau–Heckman market wage–nonmarket wage model, in the Roy model (Heckman and Honoré 1990), or in most versions of the switching regressions model. These limitations on the application of the IV method were spelled out by Heckman and Robb (1985, 1986) and later reiterated by Heckman (1995). In the switching regression context, they were discussed by Quandt (1988). Although space limitations preclude the full development of the point, IV estimation of LATE requires the same stringent behavioral assumptions.

The draft lottery number cited by AIR as a valid instrument is unlikely to satisfy (A-2) or (A-3) and thus is not likely to be a valid instrument. Consider the application of IV by Angrist (1990) that AIR discuss. The potential outcomes are earnings if persons serve in the military or if they do not. Persons who get a high number are virtually guaranteed that they are exempt from service. Those persons with a high number who nonetheless volunteer to go to the Army perceive a high gain from doing so. If those perceptions are related to the potential outcomes and are based on private information that cannot be fully predicted by $\mathbf{X}$ and $\mathbf{Z}$, then the lottery number is not a proper instrument. In addition, persons with high numbers are likely to receive more job training, because their likelihood of being drafted is reduced and firms have less likelihood of losing them. Then $\mathbf{Z}$ is an $\mathbf{X}$, and the exclusion assumption is violated. Because of the stringent nature of the required assumptions, most economists have been very cautious about using IV to identify the parameters of switching models. Sometimes, however, application of IV can be justified in the context of heterogeneous treatments. Robinson (1989), using a test proposed by Heckman and Robb (1985, 1986), demonstrated that IV methods produce appropriate estimates for estimating the "causal effect" of unions on wages; that is, the union–non-union wage differential. Robinson's evidence is surprising because it indicates that union membership is not based on unobserved components of union wage differentials not predicted by the crude $\mathbf{X}$ and $\mathbf{Z}$ available to him.

A major difference between the approach taken by AIR and that used by econometricians is that the latter go to much greater depth in justifying the behavioral assumptions that are implicit in the statistical assumptions. It is disappointing to see an entire literature in econometrics that develops explicit models designed to test and justify (A-1)–(A-3), or other identifying assumptions, ignored by AIR in their discussion of the econometrics literature.

4. IDENTIFICATION UNDER MORE GENERAL CONDITIONS FOR A VARIETY OF PARAMETERS

Heckman and Honoré (1989, 1990) presented conditions for identifiability of the full distributions of outcomes in the competing risks and Roy models. Heckman (1990) considered nonparametric identifiability in more general models. Björklund and Moffitt (1986) considered estimation of the more general models under specific distributional assumptions. Those authors demonstrate that other methods besides IV estimate behaviorally interesting parameters under more behaviorally plausible conditions. These more general models produce identification of a large array of distinct counterfactuals—a central goal of structural econometric policy evaluation—and do not focus on just one special parameter.

ADDITIONAL REFERENCES

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Bjorklund, A., and Moffitt, R. (1987), "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69, 42–49.

Falmagne, J. (1985), *Elements of Psychophysical Theory*, Oxford, U.K.: Clarendon Press.

Flinn, C., and Heckman, J. (1982), "New Methods for Analyzing Structural

Models of Labor Force Dynamics," *Journal of Econometrics*, 18, 115–168.
Gronau, R. (1974), "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.
Hansen, L., and Sargent, T. (1991), *Rational Expectations Econometrics*, Boulder, CO: Westview Press.
Heckman, J. (1974), "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 46, 695–712.
——— (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 695–712.
——— (1995), "Instrumental Variables: A Cautionary Tale," in *Essays in Honor of Sar Levitan*, eds. G. Mangum and S. Mangum, Kalamazoo, MI: W.E. Upjohn Press.
Heckman, J., and Robb, R. (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in *Drawing Inferences From Self-Selected Samples*, ed. H. Wainer, Berlin: Springer-Verlag.
Heckman, J., and Honoré, B. (1990), "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1150.
——— (1989), "The Identifiability of the Competing Risks Model," *Biometrika*, 76, 325–330.
Lucas, R., and Sargent, T. (1981), "Introduction," in *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press.
McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, pp. 105–142.

Comment

Robert A. MOFFITT

There are many unfortunate barriers to effective communication between statisticians and economists. The method of instrumental variables (IV) and associated methods for simultaneous equations and for "structural" estimation constitute one of the greatest. These methods are in the toolkit of virtually every economist and are among the most widely used techniques in the field. IV is discussed in every econometrics textbook, and three chapters of the 1984 *Handbook of Econometrics* are devoted to advanced IV and related issues, including nonlinear models (Amemiya 1984; Hausman 1984; Hsiao 1984). IV is widely regarded by economists as one of the most versatile and flexible of techniques, applicable in an enormous number of disparate applications. Yet it is scarcely used or discussed by statisticians, who often do not see the point of it all.

In this context, the attempt by Angrist, Imbens, and Rubin (AIR) to translate IV into terms that may be more understandable by statisticians must be welcomed. AIR translate IV into two frameworks familiar to statisticians. One is the well-known Rubin causal model (RCM). I find this translation to be correct and entirely appropriate, and hope that it is useful to statisticians. The other framework is the intention-to-treat (ITT) framework, with which statisticians are also quite familiar. I find this framework to have advantages as well as disadvantages. On the one hand, the noncompliance problem that is at the heart of the ITT framework is a nice illustration of the econometric problem of "endogeneity" that leads to IV estimation in economics. The notion that the difference in means between experimentals and controls should be inflated by the difference in the percentage treated in the two groups is also common to the ITT and IV frameworks. On the other hand, ITT analysis is conventionally discussed in the context of a randomized clinical trial (RCT), and AIR do so as well. This provides by necessity an obvious and convincing instrument—the experimental treatment assignment. (For another discussion of experimental assignment as an IV, see Heckman, in press.) Yet in the vast majority of work in economics, observational data are used instead, and consequently, some of the assumptions of IV stated by AIR—the random assignment assumption and the exclusion restriction—play a far more critical role in applied work than is suggested by the ITT–RCT framework.

In what follows, I make a few remarks about IV from the viewpoint of an economist, in hopes of further illuminating the interpretation and breadth of the technique. (For other recent work by economists on IV and identification in related models, see Bound, Jaeger, and Baker 1995; Manski 1990, 1994; and Staiger and Stock 1993.)

Heterogeneous Response Interpretation

The simultaneous equations model given by AIR in their equations (1)–(3) is one of the models considered in the first attempt at a comprehensive econometric treatment of the causal effects problem by Heckman and Robb (1985, 1986). That work was in turn based on the original formulation of the dummy endogenous variable model by Heckman (1978) (on which the Maddala, Bowden–Turkington, and Heckman–Robb papers cited by AIR in Section 2 are based). Although AIR find the use of unobservables in the specification of equations (1)–(3) and the assumptions surrounding it to be nonintuitive, it is important to stress that the model in those equations is nevertheless directly translatable into, and is equivalent to, the Rubin causal model (RCM) with one modification: to allow the treatment effect in equation (1) to vary across individuals; for example, $\beta_1(i)$. With that modification, $\beta_1(i) = Y_i(1) - Y_i(0)$ is the Rubin causal effect of D_i on Y_i given by AIR in Definition 2. The assumptions of linearity, additive disturbances, and other aspects of the specification in equations (1)–(3) are entirely unrestrictive in this simple a model. Thus, as

Robert A. Moffitt is Professor of Economics, Department of Economics, Johns Hopkins University, Baltimore, MD 21218. An earlier version of these remarks was presented at the meetings of the American Statistical Association, Orlando, Florida, August 13–17, 1995. The author would like to thank James Heckman, V. Joseph Hotz, and Charles Manski for comments.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Applications and Case Studies