Instrumental Variables
Response to Angrist and Imbens

James J. Heckman

I. Introduction

My 1997 JHR paper neither endorsed nor condemned the use of instrumental variables (IV) to estimate the parameters of economic models. Its main goal was to remind readers of some basic points developed in a series of papers with Richard Robb (1985, 1986). We establish that when responses to "treatment" or a policy intervention vary across persons even after conditioning on observable variables (X), a variety of mean treatment effects can be defined. When responses are homogeneous (conditional on X), all mean treatment effects are the same.

The "common effect," homogenous response, assumption underlies most currently used econometric methods used to evaluate social programs. When responses to treatment are homogenous, the case for social experiments and instrumental variables is the most persuasive. (Heckman, 1992; Heckman and Smith, 1992, 1993, 1998).

When responses to treatment vary across persons, different mean parameters can be defined that answer different economic questions. A variety of interesting policy counterfactuals can be constructed. Social experiments are better suited for recovering some of these parameters than others. In general, identification of different parameters requires different assumptions to justify application of the method of instrumental variables in any instance, or to justify the application of any econometric method. That different parameters in general require different assumptions should not be surprising. There is no universal justification of the method of instrumental variables for identifying all economic parameters.

The comment by Angrist and Imbens confuses two ideas that are fruitfully distinguished. The first idea is that every econometric estimator estimates something, provided it converges. Conditions for convergence of the IV estimator are quite weak, and involve invocation of laws of large numbers (see, for example, White, 1984).

The second, and more interesting, idea is an economic one. Does the object defined by application of the estimator answer an interesting economic question? In particu-
lar, does the method of IV recover a parameter that answers a well posed economic question?

Their discussion of my condition (C-1-b) demonstrates their confusion. That condition is sufficient for identification of the parameter Treatment on the Treated. In their notation, and mine, this parameter is $E(Y_1 - Y_0|D = 1)$. Careful readers of my *JHR* paper will note that I invoke condition (C-1-b) only for that parameter. Another condition is required for $ATE$ ($E(Y_1 - Y_0)$). That condition is (C-1-a). Yet other conditions are required to identify $LATE$. I present those conditions in my paper.

Angrist and Imbens confuse the discussion by changing the question from the one I pose and answer in my paper—what conditions are sufficient to identify Treatment on the Treated, that is, condition (C-1-b)—to the question of whether IV converges to anything, a distinctly different question.

Their “Result 1” restates a point already made in Heckman and Robb (1985, 1986) and repeated and amplified in my 1997 paper that IV does not estimate the Treatment on the Treated parameter in the general heterogenous response case unless agents do not select into the program on the basis of the unobserved heterogeneity in their response. My footnote 11 shows that IV converges to something, however, and it demonstrates that in a context of a Roy model, the parameter is exactly the $LATE$ parameter, not the Treatment on the Treated parameter. I go on to discuss what economic question $LATE$ answers. Under its identifying conditions, which are satisfied in the Roy example, it determines the gross gain to switching from the no treated state ("0") to the treated state ("1") when the policy is changed from "0" to "1"—for those induced into the military by a switch to the draft in the lottery example. It does not identify the social cost of the draft or the effect of the policy on those who do not switch, that is, the effect of military service on those who would go into the Army even if there were no draft. In general, it does not estimate Treatment on the Treated $E(Y_1 - Y_0|D = 1)$.

My *JHR* paper presents conditions under which $LATE$ answers an interesting economic question. Within the context of a utility maximizing version of the Roy model, I present conditions under which a limit form of $LATE$, defined when the policy variable $Z$ is continuous, identifies the effect of a marginal change in $Z$ on the gross gain to participants who are just indifferent or at the margin of choice between sector "0" and sector "1" or between taking treatment ("1") or not taking it ("0"). In a series of papers, Heckman and Vytlacil (1999a, b, c) and Aakvik, Heckman, and Vytlacil (1998) define the limit $LATE$ parameter as $LIV$ (local instrumental variable) and show how $LIV$ can be used to generate $LATE$, Treatment on The Treated, and the Average Treatment Effect, using an index model structure of the sort produced from a Roy model and more general economic models of discrete choice. We present bounds for these parameters when they are not identified.

Our analysis shows that a limit form of $LATE$ is the fundamental building block for constructing $ATE$, Treatment on the Treated, and $LATE$. Using the mean value function for integrals, it is possible to build up Treatment on the Treated, and the Average Treatment Effect ($ATE$) using different weighted combinations of $LATE$ parameters. If $Z$ contains one continuous regressor, one can use $LIV$ to build up $ATE$, Treatment on the Treated, and $LATE$. In a sense made precise by Heckman and Vytlacil (1999a, b, c), $LIV$ (for a model with one continuous $Z$) or $LATE$ (for a
model with only a discrete Z) are the building blocks from which Treatment on the Treated and ATE can be built up or bounded, if the support of Z is limited.

The Angrist-Imbens “Result 2” in no way casts any doubt on the validity or applicability of condition (C-1-b) to identify \( E(Y_1 - Y_0|D = 1) \) although a quick read of their comment might suggest otherwise. In fact, their result nicely illustrates the generality of (C-1-b) as an identifying condition. They present a model that satisfies that condition and IV identifies Treatment on the Treated as I claim it does. To show this, note that under their assumptions if we define \( Q = \Pr(Z = 1) \) we obtain the table of probabilities for (D, Z):

<table>
<thead>
<tr>
<th>Z</th>
<th>Pr(D = 0, Z = 0) = 1 - Q</th>
<th>Pr(D = 1, Z = 0) = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Pr(D = 0, Z = 1) = (1 - P)Q</td>
<td>Pr(D = 1, Z = 1) = PQ</td>
</tr>
</tbody>
</table>

where \( Z = 0 \) or \( Z = 1 \). If \( D = 1 \), then \( Z = 1 \). For this model, Treatment on the Treated given \( Z \) is

\[
E(Y_1 - Y_0|D = 1, Z) = \mu_1 - \mu_0 + \alpha Z
\]

while the parameter Treatment on the Treated identified by (C-1-b) is

\[
E(Y_1 - Y_0|D = 1) = \mu_1 - \mu_0 + \alpha E(Z|D = 1).
\]

But from the table we see that \( E(Z|D = 1) = 1 \) so Treatment on the Treated is

\[
E(Y_1 - Y_0|D = 1) = \mu_1 - \mu_0 + \alpha.
\]

Apply the instrumental variable condition developed in my paper. First compute the expectation of \( Y \) given \( Z \):

\[
E(Y|Z) = E(Y_0(1 - D) + Y_1(D)|Z)

= E(Y_0|D = 0, Z) \Pr(D = 0|Z) + E(Y_1|D = 1, Z) \Pr(D = 1|Z).
\]

For their model

\[
E(Y|Z) = \mu_0 + (\mu_1 + \alpha Z - \mu_0)PZ.
\]

Then for two values of the instrument, \( Z \) and \( Z' \), \( Z \neq Z' \), we obtain, under standard conditions on sampling distributions, the population parameter defined by the IV estimator:

\[
\text{Population IV} = \frac{E(Y|Z) - E(Y|Z')}{\Pr(D = 1|Z) - \Pr(D = 1|Z')}

= \frac{[(\mu_1 - \mu_0) + \alpha Z]PZ - [(\mu_1 - \mu_0) + \alpha Z']PZ'}{PZ - PZ'}

= (\mu_1 - \mu_0) + \frac{(\alpha Z)PZ - (\alpha Z')PZ'}{PZ - PZ'}.
\]
Now in their example if $Z = 1$, $Z' \neq Z$ implies $Z'$ must equal zero. Thus

$$\text{Population IV} = (\mu_1 - \mu_0) + \alpha = E(Y_1 - Y_0|D = 1)$$

as I claimed.

The confusion in their discussion of "Result 2" arises because in all of their work they assume independence of $(U_0, U_1)$ and $Z$ in defining instrumental variables, including all of their coauthored papers with Rubin. However such conditions are overly strong. For estimating mean parameters, only mean-independence of the appropriate error term is required. Condition \((C-1-b)\) does not require that either $U_1$ or $U_0$ be independent or even mean independent of $Z$. It only requires that a particular combination of the $U_0, U_1$ that constitute the error term be zero:

\[(C-1-b) \quad E(U_0 + D(U_1 - U_0 - E(U_1 - U_0|D = 1)|Z) = 0,\]

so the error term does not depend on $Z$.

Their objection to my condition is semantic and not substantive. In their example, the $Z$ dependence coming through $DU_1$ is cancelled by the $Z$ dependence coming through $E[E(U_1|D = 1)|Z]$ so $E(D, U_1|Z) = E(E(U_1|D = 1)|Z)$. It is an advantage of my more general approach that overly strong conditions like $(U_0, U_1)$ being independent of $Z$ are not required to identify parameters like Treatment on the Treated.

Parenthetically, in footnote 4 of my 1997 paper I note that if $X$ variables are introduced as direct arguments of outcome equations, the method of IV does not require that $E(U_0|X, Z) = 0$ or $E(U_1|X, Z) = 0$. It only requires that for Treatment on the Treated

$$E(U_0 + D(U_1 - U_0 - E(U_1 - U_0|D = 1), X)|X, Z) = M(X).$$

For each fixed $X$, we can difference $E(Y|Z, X)$ for distinct $Z$ values, and the $M(X)$ differences out. Heckman and Smith (1998, p. 264) provide an extensive discussion of this condition. See also Heckman, Lalonde, and Smith (1999).

These points illustrate a central theme in all of my work on program evaluation since my early papers with Robb. When responses to treatment are heterogenous, standard econometric methods have to be modified, and parameters have to be clearly defined. The case when responses are heterogeneous but agents do not participate in the program on the basis of the heterogeneity is much simpler and justifies the application of familiar methods. The method of instrumental variables is especially sensitive to the assumption of lack of information by agents. Robb and I use this sensitivity to propose an econometric test that can be used to determine if agents act on unobserved components of gains in participating in programs (1985, pp. 195–97). Robinson (1989) implements the test.

## II. The Meaning of A Causal Parameter

A major source of confusion that arises in the work of Angrist and Imbens is the definition of a causal parameter. They ignore an entire body of work in econometrics and economics that defines a causal parameter in the ceteris paribus—all other things equal (or set at specified values)—sense that economists have used since the time of Marshall (1920). Thus a ceteris paribus effect of price on
demand varies price and holds constant all other factors, both observed and unobserved.

A causal parameter is critically dependent for its definition on what is held constant. Structural economic models are causal models that recognize the constraints arising from behavioral relationships and feasibility requirements, including the possibility of interdependence among causes. It may not be possible to vary one variable without changing another one. For example if we hold the income of the consumer constant, an increase in the consumption of one good can only come at the cost of reducing the consumption of at least one other good. Angrist and Imbens’s statement that structural models are not necessarily causal models (1995, p. 433) and that the Rubin (1978) model is the only valid causal model has created confusion in the literature. Economists don’t need the Rubin (1978) model. Our structural models already improve on it by defining causal parameters that account for behavioral constraints and feasibility constraints. Structural models clarify what can and cannot be held constant when some variable or parameter is varied.¹

A structural economic model is a model of potential outcomes where the economic assumptions of the model are made explicit. Thus, for example, a structural model of labor supply writes hours of work, \( h \), as a function of wages, \( w \), unearned income, \( e \), and unobservables \( u \):

\[
h = h(w, e, u).
\]

Holding \( e \) and \( u \) fixed and varying \( w \) defines the causal and structural effect of \( w \). This is a model of potential outcomes for \( h \) as a function of \( w, e, u \). Ceteris paribus variations in \( w \) define different potential outcomes. Elsewhere (Heckman 2000), I discuss causal parameters in economics and the critical role of economic theory in defining economic causal parameters.

III. Understanding The Limitations of The Data

Angrist and Imbens combine their defense of instrumental variables with an entirely different issue: understanding what can be identified from the available data. It is a great virtue of the \textit{LATE} parameter that it makes the investigator stick to the data at hand, and separate out the aspects of an estimation that require out of sample extrapolation or theorizing from aspects of an estimation that are based on observable data.

However, this very valid and useful principle is completely distinct from the question of which particular estimator to use to impose the available information on the unrestricted data. In the standard simultaneous equations model, the reduced form contains the sample information and we obtain the system (or structural) parameters by imposing restrictions on it. A variety of different estimators can be used to impose the out-of-sample information that an analyst finds comfort in using.

More recent prototypes for this separation of conjectural from factual information are studies by Smith and Welch (1986), by Glynn, Laird, and Rubin (1986), Holland

---


Let \( f(X|D = 1) \) be the density of outcomes (for example, wages) for persons who work (for example, \( D = 1 \) corresponds to work). Suppose that we know \( \Pr(D = 1|Z) \) and hence \( \Pr(D = 0|Z) \). Missing is \( f(X|D = 0) \) (for example, wages of non-workers). In order to estimate \( E(X|Z) \), Smith and Welch (1986) use the law of iterated expectations to obtain

\[
E(X|Z) = E(X|D = 1, Z) \Pr(D = 1|Z) + E(X|D = 0, Z) \Pr(D = 0|Z).
\]

To estimate the left hand side of this expression, it is necessary to obtain information on the missing component \( E(X|D = 0, Z) \). Smith and Welch propose and implement bounds on \( E(X|D = 0, Z) \), for example,

\[
X_L < E(X|D = 0, Z) < X_U
\]

where \( X_U \) is an upper bound and \( X_L \) is a lower bound. Using this information they construct the bounds

\[
E(X|D = 1, Z) \Pr(D = 1|Z) - X_L \Pr(D = 0|Z) < E(X|Z) < E(X|D = 1, Z) \Pr(D = 1|Z) - X_U \Pr(D = 0|Z).
\]

By doing a sensitivity analysis, they produce a range of values for \( E(X|Z) \) that are explicitly dependent on the range values assumed for \( E(X|D = 0, Z) \). Glynn, Laird, and Rubin (1986) present a sensitivity analysis for distributions using Bayesian methods under a variety of different assumptions. Holland (1986) and Rosenbaum (1988, 1991, and 1995) consider more classical sensitivity analyses varying the ranges of parameters of models. The key idea in LATE, the Smith-Welch bounds and the Bayesian and classical sensitivity analyses is to clearly separate what is known from what is conjectured about the data. That is a very valuable principle that is conceptually distinct from the issue of whether or not to use IV, although LATE implements this principle.

What we can get from the available data may not always answer interesting questions. What is remarkable is that LATE sometimes answers a well posed economic question as is shown in my 1997 article and my papers with Smith (1998) and Vytlacil (1999a).

### IV. Matters of Attribution

Angrist and Imbens hint that many of the ideas in my paper are taken from other sources. It is certainly true that many of these ideas appear in Heckman and Robb (1985, 1986) and that some of these ideas have been "rediscovered" in later work. Let me comment on their principal insinuations.

2. For example, Heckman and Robb (1986, Section X, 100–104), discuss how the propensity score (the probability of selection into a program) is used differently in matching and sample selection models, and present a method for introducing the propensity score or polynomials of propensity scores as regressors to estimate selection models. This is an instance of the method of control functions introduced in their paper. Recent work by Angrist (1995) reproduces these results.
(1) I welcome their references to the work of Rubin, Peters and Belson, all of whom use the concept of Treatment on the Treated in their analyses. None of these authors presents a comprehensive discussion of alternative parameters that is featured in my papers with Robb (1985, p. 161, 1986, p. 72–77). We were unaware of these rather obscure papers in the psychology literature when we wrote our own paper and defined a variety of parameters of interest.

A more glaring omission from their list is the research of H. G. Lewis (1963) who investigated the impact of unionism on the unionized and clearly defined the parameters Treatment on the Treated and the Average Treatment Effect in both partial and general equilibrium settings. Recent joint work with Lochner and Taber (1998) builds on Lewis’ research and defines and estimates both partial and general equilibrium versions of Treatment on the Treated, ATE, and LATE. Parenthetically, we find that in a dynamic general equilibrium model of skill formation, the monotonicity assumptions of LATE are violated even when they are valid in a partial equilibrium model. We define three LATE-like effects: Direct and Reverse “LATE” and Total LATE as well as the marginal treatment effect for persons indifferent between two choices, what Vytlacil and I call the LIV parameter (Local Instrumental Variable).

(2) The structural model of schooling that I discuss in Section VII of my JHR paper owes much more to Sherwin Rosen (1977) and Becker and Chiswick (1966) than it does to David Card (1995). Rosen introduced the Wicksell (1934, pp. 178–83) schooling or “tree cutting” model into the analysis of schooling decisions. The relevance of this model to schooling choices was briefly discussed by Becker 1975 (p. 99, footnote 82). I assumed that Rosen’s work and that of Becker and Chiswick were well known. I use it to define the various treatment effect parameters. Card uses a somewhat similar model to remind readers of the point already made in Heckman and Robb (1985, p. 195–97) that instrumental variables applied to a random coefficient model of schooling, or some other variable, where the coefficient is correlated with schooling, are inconsistent, in general, except under the conditions that we specify in our work and I restate in my JHR paper. Different instruments define different parameters in the general case.

Card presents a very interesting summary of the empirical evidence on the effect of different instruments on the estimated return to schooling. However, he is not clear about what parameter he is estimating. As I read his paper, he is estimating the Average Treatment Effect. Heckman and Honoré (1990) present identification conditions for a Roy version of the schooling model when schooling is discrete valued and the coefficient on schooling is correlated with schooling. Those conditions are extended to models with more general decision structures in Heckman (1990) and Heckman and Smith (1997, 1998).

Heckman and Vytlacil (1998) discuss the random coefficient model of schooling in a systematic fashion, and analyze the various treatment effects for the Card model and the Rosen model. A major point that we make is that in Card’s model, unlike the Wicksell-Rosen model, foregone earnings are implicitly assumed not to be a cost of schooling. They are the main cost in the Rosen model. We analyze a model with both direct costs and foregone earnings costs. Ignoring the major empirical compo-

ment of the costs of schooling simplifies the econometrics but guts the economics from the model and makes it a much less interesting empirical framework for analyzing the returns to schooling.

(3) In focusing on Angrist’s military draft paper in my 1997 paper, I may have personalized the discussion too much, something I regret. I should have also discussed the first paper that used the draft lottery as an instrument. Hearst, Newman and Hulley (1986) use military draft eligibility as determined by a lottery as an instrument for the effect of military service on the post-military mortality of Vietnam veterans. Their estimator (equation 3 in their paper) can be viewed as an I.V. equation and identifies “treatment” only under the assumption that treatment effects are homogenous (at least ex ante) among people who were exempt from the draft, but volunteered anyway, and among those who were draft-eligible and were drafted. (Similarly there must be (ex ante) homogeneity in the “treatment” administered to draft-eligible persons who were inducted and persons who were draft-exempt but nonetheless chose to serve.) The clearest interpretation of their work is that they implicitly assume treatment homogeneity, a sufficient condition the validity of C-1-b.

(4) It is difficult for me to discuss the unpublished papers that I have not seen that are referred to by Angrist and Imbens in their comment. My own work with Robb on the limitations of IV methods in models with utility maximizing heterogeneous agents was done in the early 1980s. My paper with Vytlacil (1998) discusses these issues further.

References


