Instrumental Variables Before and LATEr

Toru Kitagawa

Abstract. The modern formulation of the instrumental variable methods initiated the valuable interactions between economics and statistics literatures of causal inference and fueled new innovations of the idea. It helped resolving the long-standing confusion that the statisticians used to have on the method, and encouraged the economists to rethink how to make use of instrumental variables in policy analysis.

Key words and phrases: Instrumental variables, treatment effect, treatment choice.

It is an honor to comment on Professor Imbens' paper on instrumental variables methods. The discussed paper reviews both the origin of the instrumental variables methods in econometrics and their modern formulation and interpretation based on the concept of potential outcomes originating in statistics. A unique feature of this review article is its comparative perspective. Imbens convinces us that "choice versus chance in treatment assignment" best summarizes the difference between econometrics and statistics in their traditions of identifying causal effects.

The seminal papers by Angrist, Imbens and Rubin (Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996) on the potential outcome-based formulation of the instrumental variables method are some of the few rare works that generated equally enormous influence on both econometrics and statistics communities. In the economics side, the major impacts appear in the following three aspects. First, the modern way of viewing an instrumental variable in relation to treatment noncompliance and an encouragement design widened the scope of applications of the method. Traditionally, the uses of the instrumental variables method were restricted to observational studies, and economic theories or researcher's background knowledge on the problem were playing a unique role in validating the exogeneity and exclusion restrictions of the employed instrument. Nowadays, this new encouragement design viewpoint offers another strategy for finding an instrument in a given application, and with a randomized

Toru Kitagawa is Lecturer, Department of Economics, University College London, London, WC1E 6BT, United Kingdom (e-mail: t.kitagawa@ucl.ac.uk).

initial treatment assignment, researchers can validate easily and credibly the instrument exogeneity assumption without resorting to an economic theory. Second, the concept of the local average treatment effect considerably changed the way we interpret the estimation results. We are no longer puzzled by obtaining contradicting estimation results across different instruments, and we treat them as separate and valuable pieces of information about heterogeneous causal effects. In addition, acknowledging nonidentfiability of the population average causal effect has promoted the discussion of whether the instrumental variable method should be used for the actual policy decision making and how. Third, the discovery of the importance of instrument monotonicity assumption led us to think more carefully about the subjects' causal/behavioral responses to the assigned instrument.

In what follows, I first illustrate by an example the link between the textbook linear instrumental variable model and the potential outcome framework to complement the discussion that Imbens gave in Section 6. In the second part, I review the active but unsettled discussions about usefulness of estimating the local average treatment effect, and provide briefly my personal opinion on the issue.

1. CAUSAL INTERPRETATION IN THE TEXTBOOK MODEL

The standard econometrics education introduces the instrumental variables methods in the form of, what Imbens called, the standard textbook set up,

$$(1.1) \qquad Y_i^{\text{obs}} = \beta_0 + \beta_1 X_i^{\text{obs}} + \beta_2' V_i + \varepsilon_i,$$

where Y_i^{obs} is an outcome observation of unit i, X_i^{obs} is a treatment variable of which the causal effect on the outcome is of interest, V_i is a vector of observable covariates (often called control covariates), and ε_i is an unobservable term often called as an unobserved heterogeneity of unit i. A common way to motivate the use of instrumental variables is by invalidating the least square method due to "the correlation between X_i^{obs} and ε_i." This quick but somewhat less rigorous way of motivating the instrumental variables methods often creates confusions. If equation (1.1) were specifying a regression equation or a linear projection, then the projection residual ε_i is by construction uncorrelated with X_i^{obs}, and, accordingly, the concern about endogeneity $E(X_i^{\text{obs}}\varepsilon_i) \neq 0$ would never arise. In other words, whenever instrumental variable methods are invoked, it is fundamental to understand what feature or interpretation of (1.1) distinguishes it from the statistical regression equation, and for what reason we should suspect the dependence of X_i^{obs} and ε_i.

Having a simple example would help us answer these questions. Consider a classical problem of estimation of a production function. Q denotes the quantity of a homogeneous good produced and L is the measure of labor input used (e.g., total hours worked by the employees). We do not consider control covariates for now. Assume that the production technology of firm i is given by the following function,

$$Q_i(L) = \exp(\beta_0 + \alpha_i)L^{\beta_1}, \quad 0 < \beta_1 < 1,$$

where β_0 is an unknown constant, α_i is a mean zero unobserved productivity of firm i, and β_1 is the parameter of interest assumed to be constant across firms. The specified production function leads to a log-linear equation,

$$(1.2) \qquad Y_i(x) = \beta_0 + \beta_1 x + \alpha_i,$$

where $x = \log L$ and $Y_i(x) = \log Q_i(L)$. This equation can be indeed interpreted as the causal relationship between output and input in the production process of firm i. As in equation (3.3) of the Imbens' article, $Y_i(x)$ can be interpreted as i's potential outcomes at each possible input level $x \in \mathcal{X}$. In econometrics terminology, equation (1.2) is interpreted as a *structural equation* in the sense that it can generate any counterfactual outcomes of unit i with respect to any manipulations in x. Note that the structural equation (1.2) relies only on the assumption or knowledge about the underlying causal mechanism (production function) and, so far, no considerations on how the data are generated have entered our discussion yet.

Suppose that available data of pairs of log-output and log-input of n producers, $(Y_i^{\text{obs}}, X_i^{\text{obs}})$, $i = 1, \ldots, n$, are observational, meaning that the observed input level X_i^{obs} can be seen as a "choice" made by a firm i. Following Marschak and Andrews (1944), let us model each firm's choice of X based on the following three assumptions, (1) firms are *rational*, meaning that each firm chooses its input to maximize own profit, (2) the market is under *perfect competition*, implying that every firm treat prices of the good and input (wage) as given and (3) firms have complete knowledge of their production technologies β_0, β_1 and α_i when they choose their input levels. Under these somewhat unrealistic assumptions, firm i's input choice solves the following profit maximization problem:

$$X_i^{\text{obs}} = \log L_i^{\text{obs}},$$

where

$$L_i^{\text{obs}} = \arg\max_L \{pQ_i(L) - w_i L_i\},$$

where p is the (common) price of the good, and w_i is the hourly wage given to firm i, which can vary over i, that is, the wage is determined at a localized labor market. The resulting choice X_i^{obs} is

$$(1.3) \quad X_i^{\text{obs}} = \frac{1}{1-\beta_1}\left[\beta_0 + \log\left(\frac{p\beta_1}{w_i}\right) + \alpha_i\right].$$

If we replace x with X_i^{obs} in (1.2) and notate $Y_i^{\text{obs}} = Y_i(X_i^{\text{obs}})$, we obtain

$$(1.4) \qquad Y_i^{\text{obs}} = \beta_0 + \beta_1 X_i^{\text{obs}} + \alpha_i.$$

This equation coincides with an equation of the form (1.1) without covariates. Equation (1.3) says that a more productive (higher α_i) firm chooses a larger labor input, implying that the endogeneity problem $E(X_i^{\text{obs}}\alpha_i) \neq 0$ is present. Accordingly, (1.4) must differ from the linear projection equation of Y_i^{obs} onto X_i^{obs}, and the least squares regression of Y_i^{obs} onto X_i^{obs} fails to consistently estimate β_1. Here, the keypoints are (1) there is a specific causal model (1.2) underlying (1.4), and (2) the subject's optimal "choice" based on the unobservable (to data analysts) causes correlation $E(X_i^{\text{obs}}\alpha_i) \neq 0$.

What can be a reasonable instrumental variable in the current example? A search for an instrumental variable can also be model-based. For instance, if w_i is available in data, equation (1.3) says that X_i^{obs} should be dependent on w_i, while structural equation (1.2) says w_i does not directly affect the output; accordingly, w_i satisfies the instrument relevance and the instrument

exclusion restriction. The validity of random assignment $E(w_i \alpha_i) = 0$, on the other hand, would be questionable. For instance, firms located in an urban area can be more productive (higher α_i) than those located in a rural area, and the wage level in urban area can be higher than the wage level in rural, possibly due to a higher living cost and availability of more skilled labor force. The motivation for using control covariates V_i (e.g., a demeaned indicator of whether firm i is located in an urban area or in a rural area) is to cope with potential confounders of w_i and α_i. Following the way in which Imbens treats covariates (Section 6), we assume conditional random assignment $w_i \perp \alpha_i | V_i$, and specify the dependence of α_i and V_i as

$$(1.5) \qquad \alpha_i = \beta_2 V_i + \varepsilon_i \quad \text{with } \varepsilon_i \perp V_i.$$

Here, ε_i is firm i's unobserved productivity measured relative to conditional mean $E(\alpha_i | V_i)$. Note that coefficient parameter β_2 summarizes the dependence of α_i and V_i, and we are not attaching a causal interpretation to β_2. Plugging α_i into (1.4) yields the textbook setup of the linear instrumental variable model (1.1), for which the two stage least squares procedure yields a consistent estimator for $(\beta_0, \beta_1, \beta_2)$. As is clear through this simple example, the textbook equation (1.1) can be seen as a *composite* of the causal (structural) equation (1.4) and the statistical dependence relationship (1.5).

2. POINT ESTIMATE VERSUS BOUNDS: A TREATMENT CHOICE PERSPECTIVE

The discussed paper also reviews the current debate about the meaningfulness of the complier's causal effect (Section 4.6). Imbens advocates the importance and practical values of reporting the complier's causal effect for the reason that it is the only causal estimand point-identified under the maintained assumptions. Imbens, at the same time, acknowledges that the population average causal effect is a parameter of primary interest in many contexts of causal inference, and he recommends to report also the bounds of the population average causal effect. In my opinion, Imbens' proposal is quite sensible if the main task of the data analyst is to make "scientific reporting" about the causal effects. The point-identified causal parameter for compliers and the set-identified causal parameter for the entire population reflect (partially) distinct aspects of the data distribution, and, importantly, the best we can learn from data under the maintained assumptions are only those.

The objectives of causal studies are not only for "scientific reporting," but also for assisting "decision making" of a policy maker. If the latter is a main task of the data analyst, then my personal view is that neither of the complier's causal effect estimate nor the bounds of the average causal effect should be the final output that the decision maker would find most useful. To make my argument more concrete, suppose that the decision maker's objective is to maximize the social welfare defined by the sum of individual outcomes over the target population. As in Chamberlain (2011), we suppose that he/she solves the treatment choice problem based on a posterior belief for the social welfare, that is, the decision maker is Bayesian. Since a comparison of the social welfare between the cases with and without implementation of the treatment depends only on the population average causal effect, the posterior distribution of the average causal effect obtained from her/his carefully specified prior input leads to the decision maker's optimal choice (see Chickering and Pearl, 1997 and Imbens and Rubin, 1997) for Bayesian estimation of the average causal effect). On the other hand, point estimates and inferential statements for the complier's causal effect and the bounds for the population causal effect do not directly guide formal decision-making.

The argument I just gave crucially relies on the Bayesian premise that the decision maker can fully specify a prior for the potential outcomes distributions. This may not always be the case depending on a context. Given the absence of a universal consensus on a "noninformative" prior, inability to specify a credible prior becomes a serious concern especially when the causal effect of interest is not identified, since the lack of identification makes the posterior sensitive to a choice of prior no matter how large the sample size is. One way to overcome this practical difficulty would be to follow Manski's (2000, 2005) frequentist approach based on the minimax and minimax regret decision principle, which relies only on the knowledge of the bounds of the population average causal effect.

The Bayesian approach and Manski's data-alone approach are each grounded in the two extreme schools of statistics. This means that there should certainly be a room for blending the aspects of these two approaches to complement their advantages and disadvantages. One compromising approach would be to perform a minimax or minimax-regret decision analysis with multiple priors/posteriors, namely, the Γ-minimax or Γ-minimax regret decision analysis (see,

e.g., Berger, 1985, Chapter 4). For instance, in the current context, we can consider constructing a *set of posteriors* of average causal effects by combining a single *posterior* for the identifiable parameters (causal effects for compliers, the mean of treatment outcome for always-takers, the mean of control outcome for never-takers) with a *collection of priors* of the nonidentified parameters (the mean of control outcome for always-takers and the mean of treatment outcome for never-takers). The collection of priors for the nonidentified parameters may represent the decision maker's partial or vague prior knowledge, or represent the degree of robustness that the decision maker wants to maintain in making the decision. Here, a single prior for the identified parameters would make sense in a scenario that the decision maker feels less anxious about a prior mis-specification for the identifiable parameters since he/she knows data will well update it. If the class of priors for the nonidentified parameters is not as large as the one that allows for arbitrary ones, the resulting posterior Γ-minimax treatment choice rule will not be as conservative as the Manski's data-alone minimax treatment choice rule based solely on the bounds. At the same time, unlike the standard Bayesian analysis with a single prior distribution, it can lead to a decision-making with taking into account the posterior sensitivity concern with respect to a choice of a prior for the nonidentified parameters.

REFERENCES

ANGRIST, J., IMBENS, G. and RUBIN, D. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

CHAMBERLAIN, G. (2011). Bayesian aspects of treatment choice. In *The Oxford Handbook of Bayesian Econometrics* (J. Geweke, G. Koop and H. van Dijk, eds.). Oxford Univ. Press, Oxford.

CHICKERING, D. and PEARL, J. (1997). A clinician's tool for analyzing non-compliance. *Computing Science and Statistics* **29** 424–431. MR1601275

IMBENS, G. and ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

MANSKI, C. (2000). Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *J. Econometrics* **95** 415–442.

MANSKI, C. F. (2005). *Social Choice with Partial Knowledge of Treatment Response*. Princeton Univ. Press, Princeton, NJ. MR2178946

MARSCHAK, J. and ANDREWS, W. H. JR. (1944). Random simultaneous equations and the theory of production. *Econometrica* **12** 143–205. MR0011941