“Instrumental Variables”

Thad Dunning

Department of Political Science

Yale University

Draft date: August 27, 2009

Prepared for inclusion in the *International Encyclopedia of Political Science*

1 Introduction

Confounding is a pervasive problem for drawing inferences about political causes and effects. In brief, individuals, countries, or other units are exposed to a "treatment" or "intervention" while other units are not. Differences in outcomes may reflect the effect of treatment, or they may be due to confounders—that is, variables associated with exposure to treatment and with the outcome.

Instrumental variables can be used to address the problem of confounding in both experiments and observational studies. In randomized controlled experiments, a coin flip determines which subjects are assigned to treatment, so subjects assigned to receive the treatment are, on average, just like subjects assigned to control. However, even in experiments there can be confounding, if subjects who accept the treatment are compared to those who refuse it. Analysts should therefore compare subjects randomly assigned to treatment to those randomly assigned to control. Instrumental-variables analysis may be used to estimate the effect of treatment on compliers (details follow below). In experiments, treatment assignment usually satisfies two key requirements for an instrumental variable: it is statistically independent of unobserved causes of the dependent variable, and it plausibly affects the outcome only through its effect on treatment receipt.

In observational studies, the problem of confounding is typically more severe, because units self-select into the treatment and control groups. Instrumental-variables analysis can be used to recover the effect of an "endogenous" treatment, that is, a treatment variable that is correlated with confounders. However, strong assumptions are often required, and these can be only partially validated from data. The use of instrumental variables in observational studies is discussed below, after a benchmark application to experimental data is first described.

2 Instrumental-Variables Analysis of Experiments

In experiments, subjects often fail to follow the treatment regime to which they are assigned. In a study of the effect of door-to-door canvassing on turnout, for example, voters who are assigned to receive a get-out-the-vote message may not answer the door (Gerber and Green 2000). It is misleading to compare subjects who answer the door to subjects who do not, because there may be confounding. However, treatment assignment can serve as an instrumental variable for treatment receipt, which allows estimation of the effect of treatment on compliers–that is, subjects who follow the treatment regime to which they are assigned.

An example from the health sciences helps to make the logic clear. In the 1960s, the Health Insurance Plan (HIP) clinical trial studied the effects of screening for breast cancer (Freedman 2005: 4-5, 15). About 31,000 women between the ages of 40 and 64 were invited for annual clinical visits and mammographies, which are X-rays designed to detect breast cancer. The group of women invited for screening is called the assigned-to-treatment group, or just the treatment group. In the control group, 31,000 women received the status quo health care. The invitation for screening was issued at random, so that women in the assigned-to-treatment group were just like women who were not, up to random error.

Table 1 shows death rates from breast cancer five years after the start of the trial. In the assigned-to-treatment group, 20,200 women or about two-thirds of women accepted the invitation to be screened, while one-third refused. It might seem natural to compare the women who received screening to those who refused. Yet women self-select into screening, and those who accept screening are different from those who refuse. There is an important confounder: richer and better-educated women tend to come in for screening, and while such women are less vulnerable to other diseases (see the final column of Table 1), they are more prone to breast cancer (probably because they tend to have fewer children, and child-bearing is protective against breast cancer).

The correct, experimental comparison is between women randomly invited to come in for screening—whether or not they were actually screened—and the whole control group. This "intention-to-treat" analysis, as it is called, shows a strong effect in relative terms. In the assigned-to-treatment group, there were 1.26 deaths per 1000 women, while there were 2.03 deaths per 1000 women in the control group. So the effect of assignment to screening is -0.77 deaths per 1000. However, intention-to-treat analysis likely understates the effect of screening: after all, one-third of the women in the assigned-to-treatment group were not actually screened.

What was the effect of screening on women in the treatment group who accepted screening? Instrumental-variables analysis answers this question. To begin, it is useful to think about the experimental population as comprised of two kinds of subjects: Compliers and Never Takers (Angrist, Imbens, and Rubin 1996, Freedman 2006). Here, Compliers are women who accept screening if they are assigned to treatment but are not screened if assigned to control, while Never Takers are women who are not screened, whether they are assigned to treatment or control. By looking at the control group alone, we cannot tell which is which: Never Takers look just like Compliers, since neither type of subject receives the treatment when assigned to the control group. In the treatment group, however, 20,200 or about two-thirds of women accepted screening. Because subjects are randomly assigned to treatment and control, the mix of Compliers and Never Takers should about the same in both groups. We can thus estimate that two-thirds of women in the control group are Compliers, just as in the treatment group.

Instrumental-variables analysis compares death rates of Compliers in the treatment group to death rates of Compliers in the control group. It is easy to measure the former quantity, because we observe which subjects are screened in the treatment group and can track their death rates. But what about the latter? First, note that Never Takers in the treatment group and Never Takers in the control group should have a similar incidence of death from breast cancer: after all, neither group was screened. In the treatment group, 16 women who refused screening–these are Never Takers–died from breast cancer. Thus, about 16 of the women who died from breast cancer in the control group are also Never Takers. (Here, the treatment and control groups are the same size; if they differed, we would use rates instead of numbers). Because 63 women died from breast cancer in the control group, this implies that about 63-16=47 of women who died from breast cancer in the control group were Compliers.

We can then fill in the third column of Table 1 for subjects assigned to control, dividing deaths from breast cancer by group size. The analysis implies a death rate of 1.14 deaths per 1000 women, among Compliers in the treatment group, and 2.33 deaths per 1000 among Compliers in the control group. Thus, the effect of screening on Compliers is 1.14 - 2.33=-1.19 deaths per 1000 women—a substantially larger effect than suggested by the intention-to-treat analysis. One may arrive at this same estimate by dividing the estimated intention-to-treat parameter by the fraction of the treatment group that was screened, that is,

$$\frac{-0.77}{0.65} = -1.19. \tag{1}$$

Table 1: Deaths from Breast Cancer and Other Causes (HIP study).

	Group size	Deaths from breast cancer	Death rate per 1,000 women	Deaths from other causes	Death rate from other causes, per 1,000 women
Assigned to treatment:					
Accepted Screening	20,200	23	1.14	428	21.19
Refused Screening	10,800	16	1.48	409	37.87
Total	31,000	39	1.26	837	27.00
Assigned to control:					
Would have accepted screening	20,200	47	2.33	–	–
Would have refused screening	10,800	16	1.48	–	–
Total	31,000	63	2.03	879	28.35

The table is adapted from Freedman (2005: 4, Table 1).

In the instrumental-variables estimator in equation (1), we are implicitly assuming that no women in the control group were screened. (In the 1960's, few women sought out mammography on their own). In other contexts, subjects who are assigned to the control group may seek out the treatment. With "double-crossover," the model for compliance would be extended to include Always-Takers–that is, subjects who receive treatment whether assigned to treatment or control—as well as Compliers and Never Takers. In the denominator of the instrumental-variables estimator analogous to equation (1), we would then need to subtract the fraction of the control group that was screened from the fraction of the treatment group that was screened.[1] Note that random assignment is crucial here, because it allows us to estimate the counterfactual outcomes for women in the control group who would have accepted screening, had they been assigned to control.

Some assumptions are required. For one, we must assume that there are no Defiers, or subjects who do the opposite of what they are told (Imbens and Angrist 1994, Angrist, Imbens, and Rubin 1996, Freedman 2006); in the HIP breast cancer study, Defiers are subjects who would take an exam if assigned to control but would refuse an exam if assigned to treatment.[2] Notice also that equation (1) estimates the causal effect of treatment for a specific subset of experimental subjects, namely, Compliers. When the effects of treatment are heterogenous for different subjects, this "local average treatment effect" (Imbens and Angrist 1994) may not in general be the same as the average causal effect of treatment for all subjects in the experimental population.

3 Instrumental-Variables Analysis of Observational Data

In observational studies, researchers do not apply the treatment or intervention: instead, the subjects select themselves into treatment or control groups. Selection is thus usually highly non-random, and there is typically confounding. However, under some conditions, researchers may exploit instrumental variables to recover the effect of an endogenous treatment variable. Just as in experiments, a valid instrumental variable must be independent of other causes of the dependent

[1] See Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996), and Freedman (2006: 706-709) for further discussion of the instrumental-variables estimator, and Freedman, Petitti, and Robins (2004: 72) for an application to screening for breast cancer.

[2] This "No Defiers" condition implies that the probability that each subject receives the treatment is weakly monotonic in treatment assignment (Imbens and Angrist 1994).

variable, and it must influence exposure to treatment but not influence the outcome, other than through its effect on exposure to treatment.[3]

Angrist (1990), for example, uses draft lottery numbers as an instrumental variable for military service during the Vietnam War. Understanding the effects of past military service on labor-market earnings is difficult, because people who choose to serve in the military may be different from those who do not, in ways that matter for future earnings. A key assertion in Angrist's study is that the draft number is as good as randomly assigned: whether one's draft number is high or low is therefore independent of factors that influence future earnings. With a dichotomous treatment (military service/no military service), the instrumental-variables estimator is analogous to equation (1), though the denominator should be adjusted for double crossover: some people dodge the draft, while others serve in the military even if they are not drafted. An important but reasonable assumption is that there are no Defiers, that is, residents who sign up for the military if not drafted but emigrate to Canada when their number comes up. Note that here, instrumental variables estimate the effect of treatment for a particular subset of subjects–those who serve in the military if drafted, but not otherwise. Whether this effect is informative about the effect of military service for other subjects may be a matter of opinion (Deacon 2009; Heckman and Urzua, 2009).

A second example comes from an influential study of the effect of growth on the probability of civil war in Africa (Miguel, Satyanath, and Sergenti 2004). Confounding poses a big problem in this research area, since many difficult-to-measure variables may affect both growth and the likelihood of civil war. However, year-to-year variation in rainfall may be "as-if" random (though see Sovey and Green 2009), and it may influence economic growth–that is, treatment receipt–without independently affecting the probability of civil war through other channels. If so, instrumental-variables analysis may allow estimation of the effect of economic growth on conflict, for those countries whose growth performance is shaped by variation in rainfall. This application

[3]The latter condition is sometimes called an "exclusion restriction," in reference to the exclusion of the instrumental variable from a causal equation governing the outcome.

illuminates another, distinct concern about the interpretation of instrumental-variables estimates: variation in rainfall may influence growth only in particular sectors, such as agriculture, and growth in distinct economic sectors may have different effects on the probability of conflict. Using rainfall to instrument for growth may capture such idiosyncratic rather than general effects, so caution may be advised when extrapolating results or making policy recommendations (Dunning 2008a).

Finally, Acemoglu, Johnson, and Robinson (2001), in a pathbreaking study of the effects of institutional arrangements on countries' economic performance, use colonial settler mortality rates as an instrumental variable for current institutions. These authors argue that settler mortality rates during colonial years do not affect current economic performance in former colonies, except through their effect on current institutions; they also argue that settler mortality is as good as randomly assigned, at least conditional on covariates. Since neither assumption is verifiable from the data, a combination of historical evidence and a priori reasoning must be used to try to validate, at least partially, these core assumptions. The portion of current institutions that is "explained" (in a statistical sense) by past settler mortality rates may also have idiosyncratic effects on economic growth, which could limit the generalizability of the findings.

4 Strengths and Limitations of Instrumental Variables

As the examples above suggest, instrumental variables provide an important tool, because they help to confront the problem of confounding—a first-order issue in the social sciences. Instrumental-variables regression may also be used to correct for error in the measurement of independent variables, which can pose important inferential obstacles in social-scientific research (see e.g. Sovey and Green 2009). In recent years, instrumental variables have been used to estimate causal effects in many substantive domains. Angrist and Krueger (2001) describe the evolution of the use of instrumental variables in the social sciences; for pointers on statistical technique, see Freedman (2005).

Nonetheless, the use of instrumental variables often requires strong assumptions, which can be only partially validated from data. Some empirical tests can be performed to assess the central assumption that assignment to the instrumental variable is as good as random; for instance, the instrument may be shown to be uncorrelated with pre-treatment covariates (those that are determined before the intervention). A priori reasoning and detailed knowledge of the empirical context may also play an important role. In observational studies, however, because there is often no actual randomization, the validity of as-if random assignment is often a matter of opinion; this assertion may be classified along a spectrum from "less plausible" to "more plausible" (Dunning 2008b), but it is difficult to validate the placement of any given study on such a spectrum. Sovey and Green (2009) show that the key conditions for valid instrumental-variables regression are often not defended in political science applications (though practice has improved in recent years.)

Additional issues arise in many applications, often in connection with the use of multiple regression models. For instance, concerns about the endogeneity of a single treatment variable will typically lead researchers to use instrumental-variables regression. Yet analysts typically do not discuss the possible endogeneity of other covariates in their multiple regression models. (One reason may be that the number of instruments must equal or surpass the number of endogenous variables, and good instruments are difficult to find). Furthermore, instruments that are truly random may not be strongly related to an endogenous treatment; in this case, substantial small-sample bias can arise (Bound et al. 1995).

One recommendation for practice may be to report "reduced-form" results, in addition to any other analyses. (Reduced-form is a synonym for intention-to-treat; in reduced-form regressions, the outcome is regressed directly on the instrumental variable). Another recommendation may be to report instrumental-variables regressions without covariates; with one endogenous treatment variable and one valid instrument, including covariates can be unnecessary and even be harmful (for a related discussion, see Freedman 2008a,b; though see Green 2009). The estimand should be carefully defined, and difficulties that may arise when extrapolating results to other contexts and types of subjects should be considered. In multiple-regression models, the statistical model itself must be validated, to the extent possible; with regression, the identication of causal effects depends not just on the exogeneity of instrumental variables in relation to a posited regression model but also on the validity of the underlying model itself (Dunning 2008a).

Finally, it is important to emphasize that neither of the core criteria for a valid instrumental variable—that it is statistically independent of unobserved causes of the dependent variable and that it only affects the dependent variable through its effect on the endogenous treatment—are directly testable from data. Analysts using instrumental variables should defend these assertions using evidence and reasoning, to the extent possible. Yet especially outside of the experimental context, instrumental-variables estimates should also be interpreted with an appropriate degree of caution.

References

[1] Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5):1369-1401.

[2] Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80: 313-35.

[3] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444-72.

[4] Angrist, Joshua D. and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identication: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15(4):69-85.

[5] Bound, John, David Jaeger, and Regina Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of the American Statistical Association* 90:44350.

[6] Deacon, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." NBER Working Paper No. 14690.

[7] Dunning, Thad. 2008a. "Model Specification in Instrumental-Variables Regression." *Political Analysis* 16 (3): 290-302.

[8] Dunning, Thad. 2008b. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 (2): 282-293.

[9] Freedman, David. 2005. *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

[10] Freedman, David. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691-713.

[11] Freedman, David. 2008a. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40: 180-193.

[12] Freedman, David. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2: 176-96.

[13] Freedman, David A., D.B. Petitti, and J.M. Robins. 2004. "On the Efficacy of Screening for Breast Cancer." *International Journal of Epidemiology* 33: 43-73.

[14] Gerber, Alan and Donald Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *The American Political Science Review* 94 (3): 653-663.

[15] Green, Donald. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science?" Paper prepared for the 26th Annual Society for Political Methodology Summer Conference, Yale University, July 23-25, 2009.

[16] Imbens, Guido W. and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62: 467-75.

[17] Heckman, James and Sergio Urzua. 2004. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." NBER Working Paper No. 14706.

[18] Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112(4):725-753.

[19] Sovey, Allison J. and Donald P. Green. "Instrumental-Variables Estimation in Political Science: A Readers' Guide." Paper prepared for the 26th Annual Society for Political Methodology Summer Conference, Yale University, July 23-25, 2009.