A review of instrumental variables estimation of treatment effects in the applied health sciences

Paul Grootendorst

Faculty of Pharmacy, University of Toronto
Department of Economics, McMaster University

Contact Information
paul.grootendorst@utoronto.ca
Faculty of Pharmacy, 144 College St, Toronto ON M5S 3M2
416 946 3994

Abstract

Health scientists often use observational data to estimate treatment effects when controlled experiments are not feasible. A limitation of observational research is non-random selection of subjects into different treatments, potentially leading to selection bias. The 2 commonly used solutions to this problem – covariate adjustment and fully parametric models – are limited by strong and untestable assumptions. Instrumental variables estimation can be a viable alternative. In this paper, I review examples of the application of IV in the health sciences, I show how the IV estimator works, I discuss the factors that affect its performance, I review how the interpretation of the IV estimator changes when treatment effects vary by individual, and consider the application of IV to nonlinear models.

Forthcoming in *Health Services and Outcomes Research Methodology.*

Grootendorst acknowledges support from the Premier's Research Excellence Award and the research program into the Social and Economic Dimensions of an Aging Population centred at McMaster University that is primarily funded by the Social Sciences and Humanities Research Council of Canada (SSHRC) and which has received additional support from Statistics Canada. Grootendorst thanks an anonymous referee for helpful suggestions.

Background

Much of the empirical research in the applied health sciences attempts to address questions of the sort: what is the effect of x on y? The variable y is typically a dimension of health, such as blood pressure or some other clinical endpoint; the incidence of diabetes or some other disease; health related quality of life or mortality, while x could be some health-related behavior, such as cigarette smoking; an individual's socio-economic or demographic characteristic, such as income, education or age; the use of health care, such as a new pharmaceutical drug; or exposure to an environmental toxin, such as second-hand smoke. The variable x is generically referred to as a 'treatment' and the effect of x on y a 'treatment effect'.

Much of this work relies on observational data, owing to the practical, budgetary and ethical limitations on the use of controlled experiments in this area. A widely recognized problem in observational research is that, because individuals can sometimes 'choose' different values of x (for instance, the decision to smoke or not, or the decision to use a new or old drug), it is unclear to what extent differences in y reflect differences in the level of x and to what extent differences in y reflect differences in the unobserved characteristics of those who choose different levels of x. The recent controversy over the deleterious effects of hormone replacement therapy (HRT) among post-menopausal women illustrates this attribution problem. The observational data clearly indicate that HRT and cardiovascular disease (CVD) risk are negatively correlated – HRT users tend to have better heart health than non-users. Recent experimental evidence, however, indicates that the causal impact of HRT is to *increase* the risk of heart attack and stroke. The negative correlation between HRT use and CVD risk was entirely due to the fact that generally healthy women tended to initiate HRT. (Women with more education and more income, who were typically healthy, were the ones initiating HRT so as to prevent heart disease.) In effect, HRT users had lower rates of CVD than non-users *despite* the deleterious effects of HRT.

Several solutions to this problem have been proposed, but none are entirely satisfactory. The most commonly used of these is to identify, measure and adjust for the behavioural and other factors w that are correlated with both x and y – the so-called 'covariate adjustment' approach. If ones goal is to assess the causal effect of HRT on CVD, for instance, one might adjust for various dimensions of socio-economic status and other potential 'confounding' factors correlated with HRT use that independently affect CVD risk; one could adjust for these confounding factors using either regression or matching.

Regression amounts to imposing restrictions on how w and x affect the mean of y; a common restriction is that the conditional mean of y is linear in a vector of unknown parameters α, β, γ: $E[y|x,w] = \alpha + x\beta + w'\gamma$. These unknown parameters can be estimated using ordinary least squares (OLS). The estimate of β is of primary interest – it reflects the impact of a one unit change in x on the mean of y, holding constant the influence of w. Matching compares the values of y among subjects with different levels of x but who share common values of all of the variables in w. A defect of both these techniques is that the analyst might fail to adjust for pertinent confounding variables, because they are either unknown or not readily quantifiable. Conventional regression comes with an additional liability – it requires that one correctly specify a model of the conditional mean of y and it is unclear how specification error affects ones estimate of the impact of x on y. (That being said, in some circumstances one can estimate a regression model with minimal functional form restrictions using the methods described in Li and Racine (2007).)

An alternative to covariate adjustment is to model the correlations between unobserved confounders and outcomes; these are commonly referred to as 'fully parametric' models. The leading example is Heckman's parametric sample selection model. The assumptions embedded in these models are quite restrictive: one needs to specify functional forms for both the conditional mean of y and the joint distribution of the unobserved factors affecting x and y. It is well known that results can be highly sensitive to these assumptions, yet these assumptions cannot be directly verified.

'Instrumental variables' (IV) estimation can be a useful alternative to the covariate adjustment and fully parametric approaches. IV requires one or more instruments z – variables strongly correlated with x but uncorrelated with the unobserved determinants of y. IV isolates the variation in x which is due to variation in z and assesses how this exogenous variation in x is correlated with y. The coin toss in the context of a randomized controlled trial (RCT) on fully compliant subjects is the ideal instrument – the outcome of the coin toss completely determines treatment assignment (x = treatment, control) yet does not directly affect the outcome y. The treatment effect in this case is estimated by comparing values of y among individuals randomized to different treatment groups. In an observational study assessing the effects of HRT on CVD, one might use the consumer price of HRT as an instrument, assuming that the consumer price of HRT affects its use but is uncorrelated with unobserved determinants of CVD. The latter assumption would rule out, for instance, healthier women having particularly generous drug insurance while less healthy women remaining uninsured. Alternative instruments for HRT use could include regional or time based variation in physician practice styles.

In this paper, I review examples of the application of IV in the health sciences, I show how the IV estimator works, I discuss the factors that affect its performance, I review how to interpret the IV estimator when treatment effects vary by individual, and consider the application of IV to nonlinear models.

Applications of IV estimation

Central to the use of IV estimation is the identification of good instruments. Good instruments are both *powerful* (i.e. they induce marked variation in treatments) and *valid* (i.e. they are unrelated to the unmeasured determinants of health outcomes). Here are several examples of the creative application of IV in the health sciences:

Cholera transmission

Although IV theory has been developed primarily by economists, the method originated in epidemiology. IV was used to investigate the route of cholera transmission during the London cholera epidemic of 1853-54. A scientist from that era, John Snow, hypothesized that cholera was waterborne. To test this, he could have tested whether those who drank purer water had lower risk of contracting cholera. In other words, he could have assessed the correlation between water purity (x) and cholera incidence (y). Yet, as Deaton (1997) notes, this would not have been convincing: "The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the 'poison miasmas' that were then thought to be the cause of cholera." Snow instead identified an instrument that was strongly correlated with water purity yet uncorrelated with other determinants of cholera incidence, both observed and unobserved. This instrument was the identity of the company supplying households with drinking water. At the time, Londoners received drinking water directly from the Thames River. One company, the Lambeth water company, drew water

at a point in the Thames above the main sewage discharge; another, the Southwark and Vauxhall company, took water below the discharge. Hence the instrument z was strongly correlated with water purity x. The instrument was also uncorrelated with the unobserved determinants of cholera incidence (y). According to Snow (1855, pages 74-75), the households served by the two companies were quite similar; indeed:

"the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. . . . The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity."

Is waiting for surgery hazardous?

This is a tricky question to answer. Ones position in the queue for surgery is often influenced by ones disease severity. If disease severity is imperfectly observed and controlled for, then conventional estimators of the impact of waiting time on outcomes will be biased. If more severely compromised patients are given priority access to surgery, for instance, then one might conclude that waiting is not only innocuous, it is beneficial! IV is useful if one can find good instruments for waiting time. Ho and colleagues (2000) used day of admission as an instrument for waiting time for hip fracture surgery given that hospitals perform fewer operations on weekends. Those admitted to hospital immediately prior to a weekend will therefore tend to wait longer for surgery than those admitted on other days; moreover, day of admission is likely unrelated to disease severity. Howard (2000) used the blood type of recipients of donor livers as an instrument for waiting time for liver transplants. He notes: "A person with type A blood may receive an organ from a type O donor but not a type B donor. Type O recipients are compatible only with type O donors, and type AB recipients may receive livers from any donor. Due to mismatches in the population distributions of donors' and recipients' blood types and the problems of blood type compatibility, persons with type O blood have the longest waiting times for organs and persons with type AB blood have the shortest." Organ recipient blood type appears to be a very good instrument. First, it appears to be highly predictive of waiting times for transplant surgery: holding constant patient age, sex, a battery of clinical markers and other variables, patients with type O blood waited on average 3 months longer than type AB patients. Second, it appeared to be uncorrelated with the unobserved determinants of health outcomes: after reviewing the biomedical and physiological literature, Howard could find no evidence that blood type has a direct effect on the study endpoints, graft failure at 3 months and 1 year post transplant.

Is more healthcare better?

There are marked regional variations in the healthcare services provided to Medicare beneficiaries. Fisher *et al* (2003a, b) assessed whether Medicare beneficiaries treated for hip fracture, colorectal cancer, or acute myocardial infarction (MI, heart attack) residing in high spending regions of the US – those who receive above-average quantities of inpatient and ambulatory care services – receive higher quality care, are more satisfied with their care and are healthier as a result. If not, spending on these beneficiaries could be safely reduced and deployed elsewhere. To address this question, the investigators could have assessed the correlation between patient-specific health care spending and health outcomes. The problem with this approach is that sicker patients need more health care; in

other words health care use both causes and is caused by patient health needs. To overcome this simultaneity problem, they used average beneficiary end-of-life (EOL) medical spending in the region as an instrument for individual beneficiary healthcare services use. This is a good instrument: Fisher and colleagues report that average baseline health status of study subjects was similar across regions of differing EOL spending levels (suggesting that the EOL does not directly affect health outcomes), but patients in higher-spending regions received approximately 60% more care (suggesting that EOL spending is highly predictive of the amount of care each beneficiary receives).

Effectiveness of cardiac catheterization

Newhouse and McClellan (1998) assessed the impact of cardiac catheterization (i.e. bypass surgery or angioplasty) on post MI survival rates. Comparing the health outcomes of those who do and do not get cardiac catheterization is problematic if individuals are selected for the procedure on the basis of disease severity. Instead Newhouse and McClellan used variation in distance between a patient's home and the nearest hospital with cardiac catheterization facilities as an instrument for whether the patient received cardiac catheterization after an MI. After an MI, the patient is typically rushed by ambulance to the nearest hospital, irrespective of whether or not the hospital has catheterization facilities. Hence the catheterization capability status of the hospital closest to the patient's home is a good instrument – it is a strong predictor of whether or not the patient gets the procedure and should be uncorrelated with health outcome (unless those who are more likely to suffer from severe heart attacks move to a home in close proximity to a hospital with catheterization facilities, something that the authors thought was unlikely).

Stukel *et al* (2007) also assessed the impact of cardiac catheterization on post MI survival rates using IV. In their study, they used the regional cardiac catheterization rate as an instrument for whether or not a particular patient received the procedure. This instrument was found to be highly predictive of the receipt of the procedure, more so than the differential distance instrument used by Newhouse and McClellan. One concern with this instrument, however, is that regional catheterization rates might be higher where patient needs for the procedure are greater. But Stukel and colleagues controlled for a battery of risk factors that are strongly prognostic of post-MI mortality. Hence they thought it highly unlikely that the instrument would be correlated with the residual, unobserved component of post MI mortality.

The effects of environmental exposure on future health outcomes

Jones (2007) discusses three recent studies that have followed cohorts of individuals who faced different environmental conditions in their early years to assess how nutritional and socio-economic conditions at young ages affect subsequent morbidity and mortality. Almond (2006) assessed how *in utero* fetal environment affects subsequent socio-economic status and disability. Variation in the prevalence of influenza across the United States during the flu pandemic in the fall of 1918 was used as an instrument for the *in utero* environment of those born shortly thereafter. Berg et al. (2006) assessed how socio-economic conditions during infancy affect life expectancy. They used macroeconomic conditions during infancy of those born in the 19th century Netherlands as an instrument for socioeconomic conditions. Lleras-Muney (2005) assessed how years of schooling affects life expectancy. Because both years of schooling and health are probably correlated with motivation and other latent factors, Lleras-Muney used as an instrument for years of educational attainment changes across the United States in compulsory schooling and child labour laws between 1915 and 1939. All three studies

are convincing because they used good instruments. In each case, the instrument was valid: the change in the environment exposure was due to developments, such as the onset of the flu pandemic or State policy changes, that were likely unanticipated by most individuals and hence should have affected health outcomes only through their effect on environmental exposure. Moreover, these developments preceded subsequent health outcomes. The instruments were also powerful, causing marked variation in environmental exposure.

The IV estimator

The following example illustrates how the IV estimator adjusts for the influence of confounders. Suppose that we wish to compare the effect on some continuous measure of patient health H of two types of health care, a new type of care and an existing standard. Let the variable D represent the type of care used, with $D = 1$ if the patient uses new care and $D = 0$ if the patients uses standard care. Suppose that, unbeknownst to the investigator, the value of D is assigned on the basis of patient frailty F – more frail patients are more likely to receive the new care and also have worse health. Finally, suppose that treatment D affects health H as follows:

$$H = \beta_0 + \beta_1 D + \varepsilon \qquad \text{(H1)}$$

where β_0 and β_1 are unknown parameters and the 'error' ε represents the combined influence of all determinants of H that are not explicitly modeled; one such determinant is F. Interest centers on generating consistent estimates of the treatment effect parameter β_1, the difference in effectiveness of new care and standard care. Note that I assume that treatment effectiveness is the same for everyone. Below I discuss the consequences for IV estimation when treatment effects differ by individual.

Conventional estimators of β_1 will be biased downwards. Consider, for instance, the difference in means (DIM) estimator, which is the sample average H of new care users less the sample average H of standard care users. The expected value of the DIM estimator, b_1^{dm}, is:

$$E\left[b_1^{dm}\right] = E[H|D = 1] - E[H|D = 0] \qquad (2)$$

where $E[H|D = 1]$ denotes the expectation of H for those assigned new care. The term $E[H|D = 1]$ can be evaluated by computing the expected value of the health outcomes process (H1) conditional on D:

$$E[H|D] = \beta_0 + \beta_1 D + E[\varepsilon|D] \qquad (3)$$

The expected value of H of new care users is:

$$E[H|D = 1] = \beta_0 + \beta_1 + E[\varepsilon|D = 1] \qquad (4)$$

and the expected value of H of standard care users is:

$$E[H|D = 0] = \beta_0 + E[\varepsilon|D = 0] \qquad (5)$$

Subbing (4) and (5) into (2) yields:

$$E\left[b_1^{dm}\right] = \beta_1 + E[\varepsilon|D=1] - E[\varepsilon|D=0] \quad (6)$$

It is clear that the DIM estimator will be unbiased (i.e. $E\left[b_1^{dm}\right] = \beta_1$) if and only if the expected errors are the same in both treatment groups:

$$E[\varepsilon|D=1] = E[\varepsilon|D=0] \quad (7)$$

In an RCT with fully compliant patients, random allocation of patients to treatments will ensure that (7) is satisfied. But if allocation is not controlled by the investigator, the condition might not be satisfied. If frailer patients tend to get new care and be in worse health, then $E[\varepsilon|D=1] < E[\varepsilon|D=0]$. In this case, the DIM estimator will systematically underestimate treatment effectiveness: $E\left[b_1^{dm}\right] < \beta_1$.

As I just demonstrated, condition (7) can be violated when there are confounding variables. Condition (7) can also be violated if there is 'reverse causality', i.e., if health outcomes H directly affect treatment choice D, or if there is measurement error in D.

An IV estimator for β_1 might work if D, the treatment provided to the patient, depends in part on a variable that is independent of ε. Suppose, for instance, that there are two types of physicians, labeled C (for conservative) and L (for liberal). Suppose that C physicians tend to use standard care whereas L physicians tend to use new care. The physician's practice style, described by $DocType = \{C, L\}$, is a valid instrument if 2 conditions hold. First, there needs to be pronounced differences in physician practice styles. This means that:

$$\begin{aligned} E(D|DocType) &= Prob(D=1|DocType) \times 1 + Prob(D=0|DocType) \times 0 \\ &= Prob(D=1|DocType) \end{aligned}$$

varies with different values of $DocType$. I have assumed that it does; in particular, I have assumed that:

$$Prob(D=1|DocType=L) > Prob(D=1|DocType=C) \quad (8)$$

Second, for $DocType$ to be a valid instrument, it needs to be uncorrelated with the error ε, the unmodelled determinants of the health outcome:

$$E(\varepsilon|DocType=L) = E(\varepsilon|DocType=C) \quad (9)$$

This condition implies that there are no differences in the quality of care provided by L and C-type physicians that would result in differences in patient health outcomes, nor do sicker patients gravitate selectively towards L or C physicians. A stronger condition required to estimate consistently both β_0 and β_1 is that:

$$E(\varepsilon|DocType=L) = E(\varepsilon|DocType=C) = 0 \quad (10)$$

Conditions (8) and (9) mean that $DocType$ is a valid instrument if it affects health outcomes only through its impact on the likelihood that new care is provided.

To understand how $DocType$ can be used to consistently estimate the treatment effect, take the expectation of H conditional on $DocType$:

$$E(H|DocType) = \beta_0 + \beta_1 E(D|DocType) + E(\varepsilon|DocType) \quad (11)$$

If $DocType$ is indeed a good instrument, then conditions (8) and (10) hold. Subbing conditions (8) and (10) into (11) yields:

$$E(H|DocType) = \beta_0 + \beta_1 Prob(D = 1|DocType) + 0 \quad (12)$$

Evaluating (12) under the two values of $DocType$ yields:

$$E(H|DocType = L) = \beta_0 + \beta_1 Prob(D = 1|DocType = L)$$
$$E(H|DocType = C) = \beta_0 + \beta_1 Prob(D = 1|DocType = C)$$

These two equations can be solved for β_1:

$$\beta_1 = \frac{\mathrm{E}(H|DocType = L) - \mathrm{E}(H|DocType = C)}{\mathrm{Prob}\,(D=1 \mid DocType = L) - \mathrm{Prob}\,(D=1 \mid DocType = C)} \quad (13)$$

The IV estimator is operationalized by replacing the unknown quantities by sample estimates. Hence, for example, $E(H|DocType = L)$ is replaced by the sample average health of those patients treated by L-type physicians. $Prob(D = 1|DocType = C)$ is replaced by the sample proportion of patients treated by C-type physicians who are given new care.

This formula can be adapted to handle binary outcome variables. For instance, suppose that $H_i = 1$ if subject i has an MI and $H_i = 0$ otherwise. The treatment effect, β_1, in this case, the effect of treatment type on the probability of MI, can be written:

$$\beta_1 = \frac{\mathrm{Prob}\,(H=1 \mid DocType = L) - \mathrm{Prob}\,(H=1 \mid DocType = C)}{\mathrm{Prob}\,(D=1 \mid DocType = L) - \mathrm{Prob}\,(D=1 \mid DocType = C)} \quad (13A)$$

As before, the estimator is operationalized by replacing unknown quantities with sample estimates.

The coin toss used to assign subjects into the two treatment groups in the context of a RCT is a special case of IV estimation. Suppose treatments are assigned according to process (T1):

$$D = \begin{cases} 1 \; if \; CoinToss = Heads \\ 0 \; if \; CoinToss = Tails \end{cases} \quad (T1)$$

In an RCT, the outcome of the coin toss – not $DocType$ or F – assigns subjects to treatments. According to (T1), if $CoinToss = Heads$ then the subject gets the new care ($D = 1$), and if $CoinToss = Tails$ then the subject gets standard care ($D = 0$). Then the estimator of the treatment effect is:

$$\beta_1 = \frac{\mathrm{E}(H|CoinToss = Heads) - \mathrm{E}(H|CoinToss = Tails)}{\mathrm{Prob}\,(D=1 \mid CoinToss = Heads) - \mathrm{Prob}\,(D=1 \mid CoinToss = Tails)} \quad (14)$$

If *all* subjects who get $CoinToss = Heads$ use new care, then $Prob(D = 1 \mid CoinToss = Heads) = 1$. Similarly, if all subjects who get $CoinToss = Tails$ use standard care, then $Prob(D = 1 \mid CoinToss = Tails) = 0$. In this case, the IV estimator simplifies to:

$$\beta_1 = \mathrm{E}(H|CoinToss = Heads) - \mathrm{E}(H|CoinToss = Tails) \quad (15)$$

Replacing the expected values with the sample means gives the standard DIM estimator of the treatment effect. Of course, it is entirely possible that some subjects assigned to use new care will use standard care and likewise, some subjects assigned to standard care will use new. Such non-compliance can be handled by replacing $Prob(D = 1 \mid CoinToss = Heads)$ and $Prob(D = 1 \mid CoinToss = Tails)$ with the proportions in each group who use the new care.

The generalized IV estimator

The IV estimator can be generalized to allow for multiple instruments and explicit modeling of the impact of additional determinants of H that are known to be uncorrelated with the error. The generalized IV estimator is necessary when one's instrument is a categorical variable, in which case it needs to be represented using a set of binary 'indicator' variables. Fisher *et al* (2003), for instance, used indicator variables to assign subjects to one of five quintiles of regional average end of life expenditure. Even if one uses a single binary instrument, so that the simple estimator (13) could be used, exploiting multiple sources of independent variation in D can improve the precision of the IV estimator.

Modeling the impact of additional determinants of H will reduce the error (i.e. the amount of unmodelled variation in H) and hence make it easier to satisfy the requirement that the instrument(s) be independent of the error. This is what Stukel *et al* (2007) did when they assessed the mortality outcomes of cardiac catheterization. Specifically, they used regression methods to control for a variety of patient-level factors that are predictive of mortality risk in their sample of acute MI patients. The instrument that they used – regional cardiac catheterization rates – is likely correlated with individual patients' mortality risk; but this instrument is likely uncorrelated with the portion of the mortality risk that remains after regression adjustment for the prognostic factors. Note, however, that the use of regression adjustments incurs the risk of specifying an inaccurate model of the determinants of H. (This is the same risk that one incurs when using conventional linear regression models.)

The generalized IV estimator can be implemented in two steps.[1] First, using OLS one estimates $\widehat{D}$, the predicted values from the linear regression of D on the instruments and the additional determinants of H that are explicitly controlled for. $\widehat{D}$ essentially combines the different instruments into a single summary instrument. Then one estimates (again using OLS) the linear regression of H on $\widehat{D}$ and the additional determinants of H that are explicitly modelled. The IV treatment effect estimate is the coefficient on $\widehat{D}$. The treatment D can be either binary (treated, non-treated), or continuous. The outcome H can also be binary (such as an indicator of whether the subject experienced an MI or not) or continuous. The generalized IV estimator of the effect of a treatment on a binary outcome can be quite imprecise, however. Moreover, this estimator can yield predictions that fall outside of the range 0 to 1 and hence cannot be directly interpreted as probabilities. For these reasons, it might be preferable to use an IV estimator, discussed in the section 'IV estimation of treatment effects in binary outcome models', which specifically accommodates the binary nature of the outcome.

Technical Presentation

[1] The generalized IV model can be estimated using the `ivregress` command in the statistical software program Stata (StataCorp, 2007).

In this section, I present the details of the generalized IV estimator for those who are comfortable with matrix algebra. Suppose that the error term ε in (H1) can be decomposed as

$$\varepsilon = \boldsymbol{W}'\boldsymbol{\gamma} + \upsilon$$

where $\boldsymbol{W}$ is a set of observed determinants of H, $\boldsymbol{\gamma}$ is a conformable vector of unknown parameters and υ represents the influence of the remaining latent, unmodelled determinants of H. $\boldsymbol{W}$ is assumed to be uncorrelated with υ. Then H is determined by the linear equation:

$$H = \beta_0 + \beta_1 D + \boldsymbol{W}'\boldsymbol{\gamma} + \upsilon \qquad \text{(H2)}$$

If $E[\upsilon|D = 1] = E[\upsilon|D = 0]$, then given a set of n observations on H, D and $\boldsymbol{W}$, the parameters of (H2) can be estimated using ordinary least squares (OLS). If this condition is not satisfied, then OLS is inconsistent. But if one had access to a set of instruments $\boldsymbol{Z}$ which satisfy the conditions

$$E[\upsilon|\boldsymbol{W}, \boldsymbol{Z}] = 0$$

$$\operatorname*{plim}_{n\to\infty} n^{-1}\boldsymbol{Z}'D \neq 0$$

where n is the sample size and plim denotes the probability limit operator, then one can use the generalized IV estimator. This estimator of the parameters β_0, β_1 and $\boldsymbol{\gamma}$ solves the sample moment condition:

$$\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}^*}(\boldsymbol{H} - \beta_0 + \beta_1\boldsymbol{D} + \boldsymbol{W}\boldsymbol{\gamma}) = \boldsymbol{0} \qquad (16)$$

where $\boldsymbol{X} = [\boldsymbol{1}\ \boldsymbol{D}\ \boldsymbol{W}]$ is a matrix consisting of n observations on a constant 1, the treatment indicator D and $\boldsymbol{W}$; (the ith observation of $\boldsymbol{X}$ is denoted $\boldsymbol{X}_i$); $\boldsymbol{H} - \beta_0 + \beta_1\boldsymbol{D} + \boldsymbol{W}\boldsymbol{\gamma} = \boldsymbol{\upsilon}$ is a vector consisting of n error terms; $\boldsymbol{H}$ is the vector of n observations on the health outcome, and $\boldsymbol{P}_{\boldsymbol{Z}^*}$ is the so-called projection matrix:

$$\boldsymbol{P}_{\boldsymbol{Z}^*} = \boldsymbol{Z}^*(\boldsymbol{Z}^{*'}\boldsymbol{Z}^*)^{-\boldsymbol{1}}\boldsymbol{Z}^{*'}$$

where $\boldsymbol{Z}^* = [\boldsymbol{1}\ \boldsymbol{Z}\ \boldsymbol{W}]$ is a matrix consisting of n observations on the constant, the instruments $\boldsymbol{Z}$ and the exogenous or predetermined variables $\boldsymbol{W}$. $\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}^*}$ consists of the predicted values from regressions of each of the columns in $\boldsymbol{X}$ on $\boldsymbol{Z}^*$. (Note that the predicted values from the regressions of $\boldsymbol{1}$ and $\boldsymbol{W}$ on $\boldsymbol{Z}^*$ are the observed values $\boldsymbol{1}$ and $\boldsymbol{W}$.) Hence (16) generalizes condition (10), encountered earlier, that the instruments be orthogonal to the error. Solving the sample moment condition for the unknown parameters yields the generalized IV estimator:

$$\begin{bmatrix} b_0^{iv} \\ b_1^{iv} \\ \boldsymbol{g^{iv}} \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}^*}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}^*}\boldsymbol{H} \qquad (17)$$

This estimator can be implemented in two steps. First, one estimates $\widehat{D}$, the predicted values from the regression of D on $\boldsymbol{Z}^*$. This summary instrument is orthogonal to the error. Then one estimates the regression of H on $\widehat{D}$ and $\boldsymbol{W}$. The IV treatment effect estimate is the coefficient on $\widehat{D}$.

Behaviour of the IV estimator with weak instruments and small samples

IV is not a panacea – to use it you need instruments that are both powerful and valid. Moreover, the desirable properties of the IV estimator are guaranteed to hold only as the sample size grows very large. In samples of only modest size, estimates can be highly inaccurate. In this section, I review the behaviour of the IV estimator when these requirements are not satisfied.

In general, to get precise estimates of the impact of x on y, one needs to have either large sample sizes so as to reduce sampling error (so that the atypically large errors of sample observations in one treatment group are balanced by large errors in other treatment groups) or large variation in the values of x (so that if there is a treatment effect, the resulting variation in y can be more easily distinguished from variation in y caused by sampling error). When one uses IV, one uses the portion of the variation in x due to variation in the instrument(s). When the instrument is weak, there isn't a lot of variation in x to work with and estimates will be imprecise.

The consequences of weak instruments are also apparent from equation (13). It is clear that the treatment effect parameter β_1 is in fact a ratio of two unknown quantities: the numerator is the correlation between the instrument and H, the denominator is the correlation between the instrument and D. The weak instrument problem adversely affects this denominator. When the correlation approaches zero, meaning that the instrument explains none of the variation in D, the estimator is undefined. Even when the instrument and D are highly correlated, if the sample size is small then estimates of both correlations can be imprecise, rendering the ratio of these estimated correlations highly imprecise.

The degree of imprecision of the IV estimator can be illustrated via Monte Carlo simulation. Suppose that health outcomes H are determined according to the process:

$$H = 100 + 25D - 0.50F - 0.75age + \upsilon, \upsilon \sim U[-10,10] \quad \text{(H3)}$$

where D indicates which form of care is used, F is an index of patient frailty, where a larger value of F means greater frailty, age is patient age in years, and υ is a random variable, representing idiosyncratic factors that affect health, or perhaps measurement error in H. υ can take on any integer value in the interval $[-10,10]$, each with equal probability. The shorthand way of writing this is $\upsilon \sim U[-10,10]$, where '$\sim$' means 'distributed as', 'U' means the uniform distribution, and $[-10,10]$ defines the range of values υ can assume.[2] Hence H is lower for older, more frail subjects and is 25 units higher for those using new care $(D = 1)$ compared to those using standard care $(D = 0)$.

What remains is to describe how patients are assigned to new and old care. I consider two alternatives:

Randomization with full compliance
Here D is determined by the outcome of a coin toss (T1).

Patients assigned to D on the basis of F and $DocType$.

[2] Errors drawn from $U[-10,10]$ can take on any one of 21 different integer values: -10, -9, -8, …, -1, 0, 1, …, 8, 9, 10, each occurring with equal probability (=1/21). Subjects with the same values of D, F and age can therefore have health outcomes that differ by as much as 20 units.

In this case clinical factors, not a coin toss, determine D.

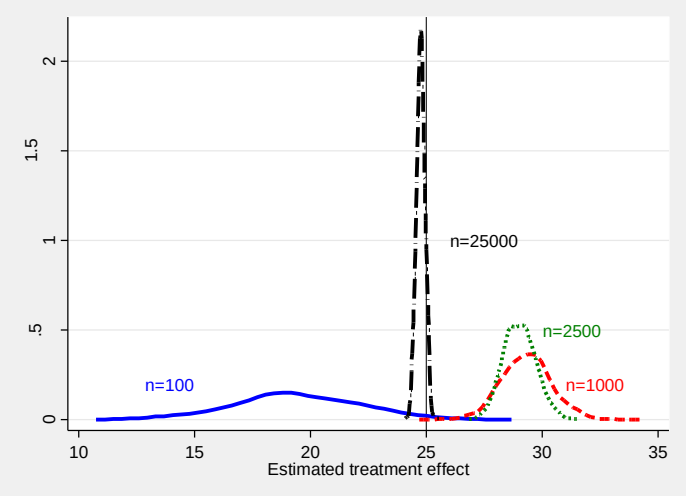
$$D = \begin{cases} 1 \; if \; Index > 0 \\ 0 \; if \; Index \leq 0 \end{cases}$$
$$Index = -40 + 0.5age + 0.5F - 20dC + \nu, \nu \sim U[-10,10] \qquad \text{(T2)}$$

New care is given only if a patient's $Index$ value exceeds some threshold level, arbitrarily set equal to zero. $Index$ is greater, the older is the patient (i.e. the greater is age), the more frail is the patient (i.e. the greater is F), and the greater is the patient-specific idiosyncratic factor (i.e. the greater is the random variable ν, where $\nu \sim U[-10,10]$). High values of ν could reflect other factors that influence treatment assignment, such as a strong patient preference for the new type of care. $Index$ is 20 points lower if the patient is treated by a C-type (conservative) doctor, identified by $dC = 1$, instead of a L-type (liberal) doctor ($dC = 0$). Hence under (T2), sicker patients and those treated by liberal doctors are more likely to receive new care.

I first demonstrate the behavior of the IV estimator as the sample size increase. To do so, I used health outcome process (H3) and treatment assignment process (T2) to generate a sample of size n observations. For each subject in my sample, I arbitrarily assigned values of dC (0 or 1), age (between 25 and 73 years), F (between 1 and 100), and randomly drawn values of υ and ν. I generated $R = 1000$ of such samples, taking independent draws of υ and ν for each sample. I used various values of $n = \{100, 1000, 2500, 25000\}$. For each of the R samples of size n observations, I used the generalized IV estimator that controls for age, and uses dC as an instrument to estimate the treatment effect parameter β_1 (which is equal to 25).

The kernel-smoothed histograms of the R treatment effect estimates are displayed in Figure 1. When $n = 100$ the sampling distribution of the estimator is rather wide (estimates varied from about 10 to 28) and is centered over 18. Increasing n to 1000 decreased the variability of estimates but did not materially improve the bias. The IV estimator is approximately unbiased when the sample size is in the order of $n = 25000$ observations.

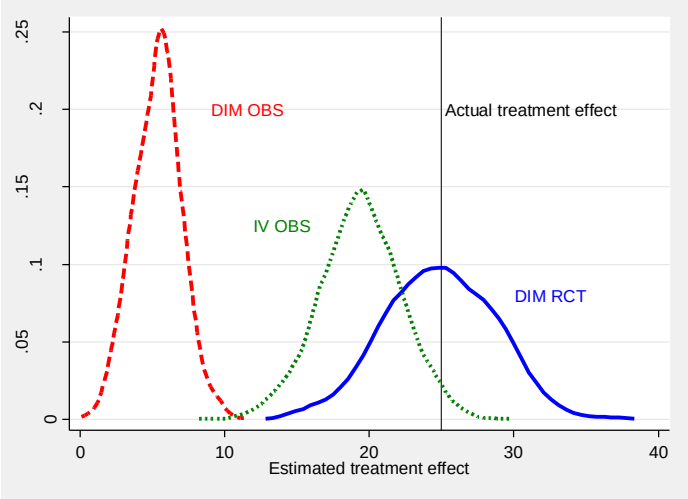
Figure 1 Histogram of 1000 treatment effect estimates generated from IV estimator using various sample sizes.



How do the other estimators compare? Using health outcome process (H3) and one of the treatment assignment processes (T1) and (T2), I generated $R = 1000$ samples of size $n = 100$ observations and, for each sample, used several different estimators to produce estimates of the treatment effect. I considered the difference in means estimator, b_1^{dm}, applied to both treatment assignment process (T2) (i.e., observational data) and treatment assignment (T1) (experimental data). I also reproduced the results for the generalized IV estimator that controls for age, and uses dC as an instrument, shown in Figure 1.

The kernel-smoothed histograms of the R treatment effect estimates are displayed in Figure 2. As expected, the sampling distribution of the DIM estimator, b_1^{dm}, estimated using the observational data (labeled as DIM OBS in the figure) is well to the left of 25, indicating that b_1^{dm} is severely downwards biased. Conversely, the DIM estimator estimated using the experimental data, DIM RCT in the figure, is centered over 25, indicating that b_1^{dm} is unbiased in this context. The IV estimator, IV OBS in the figure, is downwards biased, but not to the same degree as DIM OBS.

Figure 2 Histogram of 1000 treatment effect estimates generated from various estimators. Sample size of 100 in each case.



Note: *DIM OBS:* difference in means estimator using observational data. *IV OBS:* instrumental variables estimator using observational data (instrument is *DocType*). *DIM RCT:* difference in means estimator where treatments randomly assigned. Sample size = 100 in each case.

I next examined the properties of the IV estimator using observational data with different values of n and different degrees of correlation between the instrument and treatment assignment. First, I considered the IV estimator using a smaller sample size ($n = 50$). This case is denoted "IV n=50". Then I considered the behavior of the IV estimator where $n = 100$ as before, but where the correlation between dC and D was weakened. Specifically, I modified (T2) so that the coefficient on dC was half as large in absolute value as before:

$$Index = -40 + 0.5age + 0.5F - \mathbf{10}dC + v, v \sim U[-10,10] \quad \text{(T3)}$$

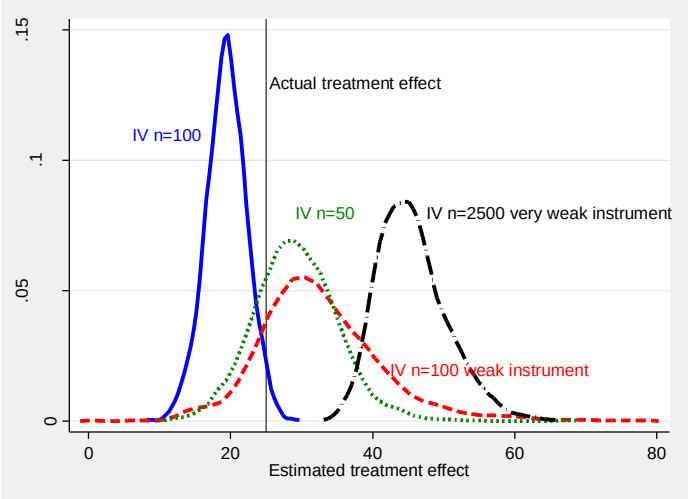
This case was denoted as "IV n=100 weak instrument". Next, I further weakened the influence of dC and D by modifying (T2) so that the coefficient on dC was even smaller:

$$Index = -40 + 0.5age + 0.5F - \mathbf{6}dC + \nu, \nu \sim U[-10,10] \quad \text{(T4)}$$

but compensated by increasing the sample size from $n = 100$ to $n = 2500$. This case was denoted as "IV n=2500 very weak instrument".

The resulting sampling distributions are displayed in Figure 3. The IV estimator using $n = 50$ is compared to the IV estimator using $n = 100$ (which appears as "IV OBS" in the previous figure). Note the smaller sample size decreases the precision of the estimator but, ironically, reduces the bias. Weakening the correlation between D and dC while using $n = 100$ resulted in a marked decrease in estimator precision – estimates ranged from less than 0 to over 80. Weakening the correlation even further while using $n = 2500$ produced an IV estimator with roughly the same precision as the "IV n=50" case, but with a marked increase in bias.

Figure 3 Histogram of 1000 treatment effect estimates generated from instrumental variables estimators.



Note: *IV n=100:* instrumental variables estimator with sample size (*n*) = 100. *IV n=50:* instrumental variables estimator with *n* = 50. *IV n=100:* instrumental variables estimator with *n* = 100 and correlation between *DocType* and *D* weakened. *IV n=2500:* instrumental variables estimator with *n* = 2500 and correlation between *DocType* and *D* weakened further. In each case instrument is *DocType.*

The take home message is that small sample sizes or weak correlation between instrument and treatment can adversely affect the performance of the IV estimator. The IV estimator can also behave poorly when an invalid instrument is used, i.e. when there is some correlation between the instrument and the unmodelled determinants of H. Indeed, Bound and colleagues (1995) demonstrate that the adverse effects of weak instruments on IV performance are exacerbated when there is even weak correlation between the instruments and the error. A final determinant of the finite sample performance of the IV estimator is the number of instruments chosen. More instruments are better, but only up to a point. As Kennedy (2003, page 175) notes, as more instruments are added, in small samples, $\widehat{D}$ becomes closer and closer to D "and so begins to introduce the bias that the IV procedure is trying to eliminate. This bias is proportional to the inverse of the *F* test statistic for testing the significance of the instrumental variables in explaining the explanatory variable for which it is to serve as an instrument."

For further discussions of weak instruments and the generalized IV estimator see Staiger and Stock (1997), Hahn and Hausman (2002), and Stock, Wright and Yogo (2002).

Are the instruments powerful and valid?

The preceding discussion suggests that it is important to assess whether the two requirements of one's instruments are satisfied.

Assumption 1: The instruments are correlated with the treatment

If they are, then the instruments should have good predictive power in the first stage regression of the generalized IV estimator, i.e. the linear regression of D on the instruments and the additional determinants of H that are explicitly controlled for. This can be tested using an *F* test of the restriction that the instruments are jointly insignificant. Small values of this *F* statistic (typically a value of less than 10) indicate a violation of Assumption 1.

Assumption 2: The instruments are uncorrelated with the error

This is technically untestable as the error is unobserved. One can get a handle on this, however, by following a practice used by experimentalists. Specifically one should ensure that the average values of the observed determinants of H are approximately equal across the different categories of the instrument (or the predicted instrument if generalized IV is used). The intuition is that if the instrument is correlated with observables that affect H, it is likely also correlated with the unobservables that enter the error term.

Assumption 2 can be formally tested if one has two or more instruments, under the assumptions that one of the instruments is valid and the treatment effect does not vary by individual. One estimates the error term using the residuals from the IV-estimated model:

$$\boldsymbol{v^{iv}} = \boldsymbol{H} - \{b_0^{iv} + b_1^{iv}\boldsymbol{D} + \boldsymbol{W}'\boldsymbol{g^{iv}}\}$$

where b_0^{iv}, b_1^{iv} and $\boldsymbol{g^{iv}}$ denote the IV estimates of the parameters in equation (H2). One then estimates a regression of $\boldsymbol{v^{iv}}$ on $(\boldsymbol{W}, \boldsymbol{Z})$. Large values of $n \times R^2$ from this regression indicates that the instruments in $\boldsymbol{Z}$ explain some of the variation in $\boldsymbol{v^{iv}}$, which is a violation of Assumption 2. If the assumption is satisfied, this test statistic is distributed χ^2 with the number of degrees of freedom equal to the number of instruments minus one. See Davidson and MacKinnon (2003) for details.

Is IV necessary?

If OLS can be used, it should be used because OLS has a much smaller variance than IV and is easier to use. OLS can be used under any of the following circumstances:

1. if one can sign the bias associated with OLS. Suppose, for instance, that OLS is known to underestimate a treatment effect. Then if the treatment effect estimate is positive despite this bias, one has good evidence that the treatment effect really is positive.

2. if the sample size is small or the instruments are only weakly correlated with the treatment D. In this case IV could be wildly inaccurate and it might be better to use OLS and accept its bias.
3. if, after controlling for observable determinants of H, the error term is not correlated with the treatment (i.e. if the expected error is same across the different treatment groups: $E[\upsilon|D = 1] = E[\upsilon|D = 0]$). If this condition is satisfied then IV and OLS should give similar estimates – they should differ only by chance. A chi-squared based test of the difference in parameter estimates can be used to verify this. An equivalent test is to estimate the residuals, $\hat{e}$, from the regression of D on $(\boldsymbol{W}, \boldsymbol{Z})$. Then one estimates the regression of H on $\boldsymbol{D}$, $\boldsymbol{W}$ and $\hat{e}$. If IV and OLS give similar estimates, the variable $\hat{e}$ will not be statistically different from zero. Johnson and DiNardo (1997, page 339) provide the intuition behind this test. The regression of D on $(\boldsymbol{W}, \boldsymbol{Z})$ splits the variation in D into two parts. One part, the predicted values from this regression (each of which is a linear combination of the variables in $(\boldsymbol{W}, \boldsymbol{Z})$), is uncorrelated with υ, assuming that the instruments are valid. The other part, the residuals, is uncorrelated with υ if the condition is satisfied. If the condition is not satisfied, then these residuals will explain successfully some of the variation in υ, (or, equivalently, some of the variation in H that remains after conditioning on $(\boldsymbol{W}, \boldsymbol{Z})$) and IV estimation may be warranted.

The reduced form model

Even if the IV estimator performs poorly, one can learn about the treatment effect by estimating via OLS what is known as the 'reduced form' model – the linear regression of the health outcome H on the instruments and any health determinants that are explicitly modeled (Angrist and Krueger 2001). The estimated effect of the instrument on H (i.e. the 'reduced form' estimate or RFE) represents the 'net' effect of the instrument on H. As I demonstrate formally below, RFE is the product of the effect of the instrument on treatment choice D (which reflects the strength of the instrument or IS) and the effect of D on H (i.e. the treatment effect or TE). Algebraically, this can be expressed as

$$RFE = IS \times TE \tag{18}$$

If RFE is close to 0, then either IS is close to 0 (the instruments are weak) or TE is close to 0 (the treatment effect is negligible) or both of these are true. If one can rule out the former using the results of the first stage regression of the generalized IV estimator, then there is evidence that TE is 0 – new treatment does not work better than the standard. If RFE is appreciably different from 0, then RFE, combined with the estimates of IS from the first stage regression of the impact of the instrument on D, can be used to deduce the magnitude and direction of the treatment effect. Specifically, rearranging (18), TE can be expressed as the ratio of RFE to IS:

$$TE = \frac{RFE}{IS} \tag{19}$$

The reduced form model is analogous to 'intention to treat' analysis of RCT data, wherein one compares the average H of subjects grouped by the treatment type that they were randomly assigned to (i.e. the value of the instrument), irrespective of the treatment actually used. Equation (19) suggests that the treatment effect can be derived from intention to treat estimates (RFE) and estimates of IS. Indeed, it is instructive to note that (19) simply restates equation (13).

As an example Evans and Ringel (1999) assess the effect of state cigarette taxes on infant low birth weight; infant birth weight is a strong predictor of infant mortality and, among surviving infants, health in later years. Cigarette taxes, of course, do not affect birth weight directly; they affect birth weight only through their impact on maternal smoking. Hence Evans and Ringel could have used cigarette taxes as an instrument for maternal smoking participation in a model of infant low birth weight status. This approach is problematic, however, for several reasons. One reason is that binary outcome models can be imprecisely estimated using IV. Instead, they estimated using probit regression the impact of cigarette taxes on infant low birth weight status (the reduced form estimate) and divided this by the estimated impact of cigarette taxes on maternal smoking participation (the first stage model) to derive an estimate of the impact of maternal smoking participation on infant low birth weight status.

To illustrate more formally, suppose that H is determined by the equation:

$$H = f(D, \boldsymbol{W}, \upsilon) \tag{18}$$

where, as before, $\boldsymbol{W}$ are the known exogenous health determinants and υ is a variable reflecting the influence of all other health determinants. And suppose that treatment assignment D is determined by the equation:

$$D = g(\boldsymbol{Z}, \boldsymbol{W}, \nu) \tag{19}$$

where $\boldsymbol{Z}$ are the instruments, and ν is a variable reflecting the influence of all other determinants of treatment choice. The reduced form equation is derived by substituting (19) into (18):

$$H = f(g(\boldsymbol{Z}, \boldsymbol{W}, \nu), \boldsymbol{W}, \upsilon) \tag{20}$$

The net effect of instrument Z_j on H is found by differentiating (20) w.r.t. Z_j:

$$\frac{\partial H}{\partial Z_j} = \frac{\partial f(D, \boldsymbol{W}, \upsilon)}{\partial D} \frac{\partial g(\boldsymbol{Z}, \boldsymbol{W}, \nu)}{\partial Z_j}$$

To learn about $\frac{\partial f(D, \boldsymbol{W}, \upsilon)}{\partial D}$, one can estimate $\frac{\partial g(\boldsymbol{Z}, \boldsymbol{W}, \nu)}{\partial Z_j}$ (from the first stage regression) and $\frac{\partial H}{\partial Z_j}$ from the reduced form regression model. The reduced form regression model:

$$H = \alpha_0 + \boldsymbol{Z}'\boldsymbol{\alpha_1} + \boldsymbol{W}'\boldsymbol{\alpha_2} + \omega$$

is an approximation to equation (20). The α's are unknown parameters that can be estimated using OLS and ω is the composite error term derived from ν and υ. The estimates of $\boldsymbol{\alpha_1}$ are the reduced form estimates.

IV estimation of variable treatment effects

The effect of some treatment x on a health outcome y might vary between individuals. For instance, some medications are less effective among 'poor metabolizers'. Some individuals apparently do not seem to suffer any adverse consequences from cigarette smoking, while others do. A recent literature has analyzed the properties of the IV estimator when treatment effects vary by individual and

individuals choose the best treatment option using private information (information not observed by the researcher).

An illustrative example is Newhouse and McClellan's analysis of the impact of post-MI cardiac revascularization on mortality rates. Recall that they used as an instrument the 'differential distance', the additional distance (if any) between the hospital closest to the patient's residence and a hospital with revascularization capacity. While this instrument was highly correlated with the receipt of revascularization, it was the not the only determining factor. Another, unobserved, factor, the patient's suitability for revascularization, was also likely important. Indeed patients who were ill suited for revascularization would almost certainly not undergo the procedure, no matter how close they lived to a hospital which could perform the procedure. Other patients who were ideal candidates for the procedure would eventually likely receive it, again, irrespective of their proximity to a hospital. So the IV estimator in this case tells us about the impact of revascularization on the subset of patients occupying the middle ground between being ideal candidates and being ill-suited for the procedure. For such patients, the effectiveness of the procedure is unclear and factors such as geographic proximity to a revascularization facility could be deciding factors in treatment decisions. Note that in this case, the IV treatment effect estimate underestimates the effectiveness of treatment for ideal candidates and overestimates the effectiveness of the treatment among those ill suited for the procedure.

To obtain more precise results about the behavior of the IV treatment effect estimator when treatment effects vary, let us modify the health outcome model (H1) slightly:

$$H_i = \beta_0 + \beta_{1i} D_i + \varepsilon_i$$

where i indexes subjects. Hence β_{1i} reflects the treatment effect specific to subject i. Imbens and Angrist (1994) demonstrate that the IV estimator converges to a weighted average of treatment effects where the weights are largest for subjects who vary their treatment choice D_i by the greatest degree in response to changes in the instruments. In particular, following the nomenclature of Auld (2006), if the instrument Z takes G different values, then under fairly general conditions the IV estimate of 'the' causal effect of D on H using Z as an instrument converges to:

$$b_1^{iv} \rightarrow \sum_{g=1}^{G} \lambda_g \beta_{1g}$$

where β_{1g} is the average effect of D on H in subpopulation g and the λ_g's are weights that depend on how much D varies with Z in subpopulation g. Imbens and Angrist call b_1^{iv} the 'local average treatment effect' (LATE). Recalling our previous example, one could imagine there being just two values of g: those who live near $(g = 1)$ and those who live far away $(g = 2)$ from a catheterization hospital. In both groups, those whose catheterization treatment decision does not depend on distance will contribute nothing to the treatment effect estimate. Hence the IV estimate reflects the average of the treatment effects in the remaining patients – those whose catheterization treatment decision depends on distance.

The lesson is that when treatment effects vary, IV estimates will tend to reflect the treatment effects of those whose treatment decision varies the most with variation in the instrument. One implication is that the interpretation of the IV estimator can depend on the instrument used. Two analysts, each using valid instruments, can legitimately produce different LATE estimates. To illustrate, suppose that catheterization treatments were assigned using a coin toss, not differential distance. Hence there would

again be two values of g: those randomized to receive catheterization treatment ($g = 1$) and those randomized to standard care ($g = 2$). If subjects were compliant, then the mix of patient types should be the same in both groups, implying that the weights λ_g and the average treatment effect β_{1g} should be the same in both groups as well. Moreover because each patient's treatment alllocation is equally dependent on the outcome of the coin toss, each patient will have an identical weight and the IV estimator will converge to a simple average of the β_{1i}. Note that the interpretation of the IV estimator in this context is different than the interpretation when distance was used as an instrument. Another implication of the LATE framework is that when several instruments are used, it can be difficult to understand whose treatment decisions are being most affected by variation in the instruments (Heckman 1997).

IV estimation of treatment effects in nonlinear models

The models consider so far have been linear in parameters. Some models, however, are nonlinear. Of these nonlinear models, some can be rendered linear via a suitable transformation. For instance, if health outcomes are determined by the process:

$$H = e^{\beta_0} e^{\beta_1 D} w^{\beta_2} e^{\upsilon} \tag{H4}$$

Then, as Davidson and MacKinnon (2003, page 22) note, the model can be rendered linear in the parameters by taking the logarithm of both sides:

$$\ln H = \beta_0 + \beta_1 D + \beta_2 \ln w + \upsilon \tag{H5}$$

IV estimation of intrinsically nonlinear models can be handled using nonlinear IV estimation. Nonlinear IV estimation is suitable when one can formulate one's model in the form of a nonlinear regression:

$$H_i = \boldsymbol{x}_i(\boldsymbol{\beta}) + \varepsilon_i \tag{21}$$

where $\boldsymbol{x}_i(\boldsymbol{\beta})$ is a nonlinear regression function that depends on $\boldsymbol{\beta}$, a vector of K unknown parameters, the treatment indicator D_i and any other covariates included in the model. As before, ε_i represents the influence of all other determinants of H_i that are not explicitly modeled, some of which may be associated with D_i.

This framework can accommodate a variety of models. Suppose, for example, that H_i is a count variable; perhaps H_i is the number of chronic health problems afflicting subject i. The Poisson model of H_i can be written as:

$$H_i = \exp(\beta_0 + \beta_1 D_i + \boldsymbol{W_i}'\boldsymbol{\gamma}) + \varepsilon_i \tag{22}$$

Conversely H_i might be a binary outcome; it could be the case that $H_i = 1$ if subject i has high blood pressure and $H_i = 0$ otherwise. The logit model of the probability that $H_i = 1$ can be written as:[3]

[3] As Davidson and MacKinnon (2003, page 456, 476) note, one could improve estimator precision by dividing observations on H_i and $\boldsymbol{x}_i(\boldsymbol{\beta})$ by the square root of the observation's error variance. The error variance in the Poisson case is equal to its conditional mean, while the error variance in the logit case is simply the variance of a Bernoulli distributed random variable: $p_i(1 - p_i)$ where $p_i = \frac{\exp(\beta_0 + \beta_1 D_i + \boldsymbol{W_i}'\boldsymbol{\gamma})}{1 + \exp(\beta_0 + \beta_1 D_i + \boldsymbol{W_i}'\boldsymbol{\gamma})}$.

$$H_i = \frac{\exp(\beta_0+\beta_1 D_i+\boldsymbol{W_i}'\boldsymbol{\gamma})}{1+\exp(\beta_0+\beta_1 D_i+\boldsymbol{W_i}'\boldsymbol{\gamma})} + \varepsilon_i \quad (23)$$

Nonlinear IV models are somewhat controversial. Terza (2006) is skeptical about the realism of models of the form (21) because they treat observed and latent determinants of health outcomes asymmetrically. While observed determinants are modeled using the nonlinear function $\boldsymbol{x}_i(\boldsymbol{\beta})$, latent determinants are relegated to the additive error term. Hence the marginal effects of observed and unobserved covariates on H can be quite different, with no apparent justification. The one exception is the exponential model (22); Terza demonstrates that this model treats observed and unobserved health determinants symmetrically.

Consistent estimation of the parameters of nonlinear regression models requires that $\widehat{\boldsymbol{\beta}}$, the estimated values of $\boldsymbol{\beta}$, satisfy the following 'moment conditions':

$$\boldsymbol{X}(\widehat{\boldsymbol{\beta}})'\left(\boldsymbol{H} - \boldsymbol{x}(\widehat{\boldsymbol{\beta}})\right) = \boldsymbol{0} \quad (24)$$

where $\boldsymbol{X}(\boldsymbol{\beta}) \equiv \frac{\partial \boldsymbol{x}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is the matrix of n observations on the K partial derivatives of the regression function with respect to each of the K parameters in $\boldsymbol{\beta}$. These moment conditions are the generalization to the nonlinear regression context of the condition in the linear context that the errors be independent of the treatment. For instance, if the ith observation on $\boldsymbol{x}_i(\boldsymbol{\beta})$ is

$$\boldsymbol{x}_i(\boldsymbol{\beta}) = \beta_0 + \beta_1 D_i$$

Then

$$\boldsymbol{X}_i(\boldsymbol{\beta})' \equiv \left(\frac{\partial \boldsymbol{x}_i(\boldsymbol{\beta})}{\partial \beta_0} \quad \frac{\partial \boldsymbol{x}_i(\boldsymbol{\beta})}{\partial \beta_1}\right) = (1 \quad D_i)$$

Hence (24) would require that D and ε be orthogonal (i.e. the errors do not vary with values of D). When the errors are not orthogonal to $\boldsymbol{X}_i(\boldsymbol{\beta})$ then one can use nonlinear IV provided that one has a set of instruments $\boldsymbol{Z}$ that satisfy the condition that:

$$\boldsymbol{X}(\widehat{\boldsymbol{\beta}})'\boldsymbol{P}_{\boldsymbol{Z}^*}\left(\boldsymbol{H} - \boldsymbol{x}(\widehat{\boldsymbol{\beta}})\right) = \boldsymbol{0} \quad (25)$$

where $\boldsymbol{P}_{\boldsymbol{Z}^*}$ was previously defined and $\boldsymbol{X}(\widehat{\boldsymbol{\beta}})'\boldsymbol{P}_{\boldsymbol{Z}^*}$ are the predicted values from regressions of the columns of $\boldsymbol{X}(\widehat{\boldsymbol{\beta}})$ on $\boldsymbol{Z}^*$. Hence condition (22) states that the summary instruments be independent of the error terms. The nonlinear IV estimator is the estimate of $\boldsymbol{\beta}$ that solves (25); this estimator is equivalent to the estimate of $\boldsymbol{\beta}$ that minimizes the criterion function:

$$Q(\boldsymbol{\beta}, \boldsymbol{H}) = \left(\boldsymbol{H} - \boldsymbol{x}(\boldsymbol{\beta})\right)'\boldsymbol{P}_{\boldsymbol{Z}^*}\left(\boldsymbol{H} - \boldsymbol{x}(\boldsymbol{\beta})\right) \quad (26)$$

IV estimation of treatment effects in binary outcome models

How can one use IV to estimate treatment effects on binary health outcomes? If there is a binary treatment and one binary instrument, then one can use equation (13A). In more general settings, with multiple instruments, and explicit controls for observed health determinants, one has the following options:

1. One could model the binary outcome using logit or probit regression and use the nonlinear IV estimator if one can find instruments satisfying (25). Note that this estimator is sometimes misused. Recall that IV estimation of the parameters of the linear model can be performed in two steps, wherein D is replaced by $\widehat{D}$, the predicted values from the regression of D on $\boldsymbol{Z}^*$, and this modified model estimated by OLS as per usual. When the model is nonlinear, IV estimates minimize the criterion function (26); replacing D with $\widehat{D}$ and estimating the modified model by nonlinear least squares will not yield consistent estimates (Davidson and MacKinnon, 1993, page 225; Gail *et al* 1984). Hence the methods used in Howard (2000) are questionable. The author estimated a probit regression of waiting time for organ transplant on the probability of graft failure. In this regression, data on actual waiting time was replaced with the predicted values from a regression of waiting time on indicators of recipient blood type and various predictors of graft success. This is technically incorrect.

2. Another option would be to estimate the reduced form model and the first stage model and, using these estimates, infer the treatment effect. For instance, Howard could have estimated the probit model for graft success on the indicators of recipient blood type and various predictors of graft success. This estimate, combined with an estimate of the impact of recipient blood type on transplant wait times, could have been used to infer the impact of waiting on the probability of graft success.

3. The generalized IV estimator could be applied to binary outcomes, but as was discussed earlier, this method can yield imprecise estimates and nonsensical predictions.

4. A 'fully parametric' model could be used. These methods require that the analyst specify the joint distribution of the unobserved determinants of treatment choice and the likelihood that the binary outcome is equal to one. The Stata command `ivprobit` implements this method. Another approach, useful if the effect of a binary treatment varies across individuals and individuals choose the best treatment option using private information, is to implement the methods used by Auld (2005). This approach requires that one specify the joint distribution for three unobserved factors: the determinants of the propensity to choose one treatment over the other, the determinants of outcomes when the new treatment is chosen and the determinants of outcomes when the standard treatment is chosen.

Concluding Remarks

Instrumental variables estimation can be a useful alternative to conventional covariate adjustment approaches. Finding good instruments, however, is not easy. Successful application of IV requires either experimental variation in treatment assignment or a source of quasi-experimental variation that is incidental to the outcome being analysed. Furthermore, the desirable properties of IV are guaranteed to hold only as the sample size grows very large. In samples of modest size, IV estimates can be wildly

inaccurate if instruments have only a modest effect on treatment or if there is even a weak correlation between instrument and the outcome being modeled. Finally, IV estimation when treatment effects are heterogenous requires careful consideration of the subjects whose treatment status is affected by variation in the instruments. IV reveals nothing about treatment effectiveness among subjects whose treatment status is non-responsive to variation in the instrument.

References

Almond D. Is the 1918 influenza pandemic over? Long term effects of in utero influence influenza exposure in the post 1940 US. *Journal of Political Economy* 2006; 114:672-712.

Angrist J, Krueger A. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives* 2001; 15(4):69-85.

Auld MC. Causal effect of early initiation on adolescent smoking patterns. *Canadian Journal of Economics* 2005; 38:709–734.

Auld MC. Using observational data to identify the effects of health-related behavior. In Andrew Jones (ed) *Elgar Companion to Health Economics*, Elgar, 2006.

Berg GJVD, Lindeboom M, Portrait F. Economic conditions early in life and individual mortality. *American Economic Review* 2006; 96:290-302.

Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 1995; 90:443– 450.

Davidson R, MacKinnon JG. *Econometric Theory and Methods*. New York: Oxford University Press, 2004.

Davidson R, MacKinnon JG. *Estimation and Inference in Econometrics*. New York: Oxford University Press, 1993.

Deaton A. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press, 1997

Evans WN, Ringel JS. Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics* 1999; 72:135–54.

Fisher ES, Wennberg D, Stukel T, Gottlieb D, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care. *Annals of Internal Medicine* 2003a; 138(4):283-287.

Fisher ES, Wennberg D, Stukel T, Gottlieb D, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. *Annals of Internal Medicine* 2003b; 138(4):288-299.

Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; 71:431-444

Hahn J, Hausman J. A new specification test for the validity of instrumental variables. *Econometrica* 2002; 70(1):163–189.

Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluation. *The Journal of Human Resources* 1997; 32: 441-462.

Ho V, Hamilton B, Roos L. Multiple approaches to assessing the effects of delays for hip fracture patients in the U.S. and Canada. *Health Services Research* 2000; 34:1499-1518.

Howard D. The impact of waiting time on liver transplant outcomes. *Health Services Research* 2000; 35: 1117–1134.

Imbens G, Angrist J. Identification and estimation of local average treatment effects. *Econometrica* 1994; 62(2):467-76.

Johnson J, DiNardo J. *Econometric Methods*, 4th edition. McGraw Hill, 1997.

Jones AM. Panel data methods and applications to health economics. *University of York Health Econometrics and Data Group Working Paper* 07/18, 2007.

Kennedy P. *A Guide to Econometrics*, 5th edition. Cambridge MA: MIT Press, 2003.

Li Q, Racine JS. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

Lleras-Muney A. The relationship between education and adult mortality in the United States. *Review of Economic Studies* 2005; 72:189-221.

Newhouse J, McClellan M. Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health* 1998; 19:17-34.

Snow J. *On the Mode of Communication of Cholera*. London: Churchill, 1855. [Reprinted (1965) by Hafner, New York.]

Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997; 65:557–586.

StataCorp. *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP, 2007.

Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 2002; 20(4): 518-529.

Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Journal of the American Medical Association* 2007;297:278-285.

Terza JV. Estimation of policy effects using parametric nonlinear models: a contextual critique of the generalized method of moments. *Journal of Health Services and Outcomes Research Methodology* 2006; 6:177-198.