THE ECONOMIC RETURNS TO EDUCATION: FINITE-SAMPLE PROPERTIES OF AN INSTRUMENTAL VARIABLE ESTIMATOR

GÁBOR KÉZDI[1]

This paper evaluates the instrumental variables measurement of the causal effect of education on earnings, with special focus on finite sample issues. A simple well-known theoretical model is presented and the inconsistency of the reduced-form estimator is established. The problem of weak instruments is examined in details and multiple remedies are considered. The relevant issues are illustrated through a particular example from a published paper, including a simulation exercise to inspect weak instruments problems and evaluate the performance of alternative estimators.

KEYWORDS: Returns to education; Instrumental variables; Weak instruments.

Instrumental variables models are very popular in empirical economics for estimating causal effects on observational data.[2] In their clearest form, causal effects can be stated in the framework of thought experiments. Because of nonrandom assignment, however, the non-experimental nature of virtually all economic data makes measurement of the thought experiments difficult. Simple reduced-form models like ordinary least squares (OLS) can be thought of as generalized versions of comparing means in different groups. The problem is that self-selection into those groups is typically not random, and therefore simple between-group comparisons do not measure the intended causal effects. Economics models often help capturing the non-randomness of the assignment and finding the direction of the resulting bias. Instrumental variables (IV) models offer more than that: under the necessary assumptions, IV results can be interpreted as estimates of the causal relationship. Problem is that the required assumptions are quite restrictive, and their validity is often difficult to assess. For accessible reviews of the IV and 'natural experiment' estimators, see *Meyer* (1995), *Angrist, Imbens* and *Rubin* (1996) with the comments and *Heckman* (1997).

The purpose of this paper is to evaluate the IV measurement of an extensively researched question, the causal effect of education on earnings, also known as the (private) economic returns to schooling. It is a good example that there is a simple economic

[1] PhD candidate, Department of Economics, University of Michigan.

[2] The topic of this paper was suggested by *John Bound* and *Jinyong Hahn*. I also thank *László Hunyadi* and *Gábor Körösi* for their helpful comments. All the remaining errors are mine.

model behind the self-selection story but the estimation raises quite a few econometric problems. *Willis* (1986) offers a thorough treatment of the model, and *Willis'* (1986) and *Card's* (1999) works are comprehensive surveys of the empirical literature. In this paper, I illustrate the problems through a specific study by *Harmon* and *Walker* (1995).

As we will see, *Harmon* and *Walker* find OLS to be biased downward. This result is not uncommon: most IV estimates reported by *Card* (1999) that use of similar instruments estimate negative or zero bias. At the same time the conventional theory of educational choice implies that OLS should be biased upwards. Either the theory is wrong or the estimation strategy is flawed. A third possibility might be measurement error in reported education level, which could induce a downward bias. It turns out, however, that the required error is an order of magnitude larger than what is established in the literature (The paper will show that in its example it has to be well over 50 per cent of the total variation.) *Card* (1999) considers a richer model than *Willis* (1986) that has predictions broad enough to incorporate some negative bias of the OLS. Harmon and Walker's estimates, similar to most estimates based on compulsory schooling instruments, are nevertheless a lot larger than all other kind of IV estimates. I am therefore sceptical about their results: I try to find out what could be wrong with the particular strategy they followed.

Among the possible problems, I will focus mainly on the finite sample properties of the estimator. In a controversial study, *Angrist* and *Krueger* (1991) had used an IV strategy similar to Harmon and Walker's to estimate returns to schooling. Their estimates also seem to suggest that OLS is not biased upward. However, *Bound, Jaeger and Baker* (1995) have shown that the instruments they used were weak and thus the results may not tell anything meaningful about the true population relationship of OLS and IV. The fact that Angrist and Krueger's IV was estimated on a very large sample but still suffers from finite-sample problems was suprising to many. The problem of 'weak instruments' was known to econometricians for quite a long time but it was typically ignored by practitioners who thought that large samples are immune to it. *Bound et al*. (1995), *Staiger* and *Stock* (1997) and others have convincingly shown, however, that weak instruments can be a problem in seemingly large samples, too. On the other hand, there exist modified IV estimators that are asymptotically equivalent to the conventional ones but have superior finite-sample properties. I will show the basic intuition behind the problem and consider a few of the alternatives.

The objective of this paper is primarily methodological. I would like to draw attention to the fact that IV can be a powerful strategy if supported by economic theory, but one should not ignore the econometric problems. In particular, the possible weakness of the instruments should be taken seriously. I would also like to show how one can detect the problem and that there are possible remedies. The causal effect of education on earnings serves a paradigmatic example, and my analysis of the *Harmon* and *Walker* (1995) study is intended to be an illustrative exercise. As it turns out, their instruments are not weak, and therefore their results are robust to finite-sample corrections. However, their strategy can be questioned on other grounds. Unfortunately, there is not enough available evidence to address those, but they suggest that it may be a little early to bury the conventional model of educational choice.

The remainder of the paper can be divided into five sections. First, it presents a relatively simple version of the theoretical model. The second part discusses the estimation

of the return to schooling. The third section focuses on the finite sample properties of the IV estimator and introduces some alternatives. The fourth discusses *Harmon* and *Walker's* (1995) study and presents results of a Monte-Carlo simulation designed for capturing the finite-sample bias of their estimator. The last part concludes.

1. THE THOUGHT EXPERIMENT

To fix ideas, consider the following thought experiment. Take an individual, assign her a random level of education, and then measure her lifetime earnings. Then assign her a different level of education and measure her lifetime earnings again. The difference of the two earning levels represents the causal effect of education. Repeating the experiment enough times on enough randomly chosen individuals, one can get a good estimate for the average causal effect of education on earnings.

Obviously, this experiment is impossible to carry out. More important is that, as we will see, we can't capture anything close to it in observational data. Therefore, the causal effect of a randomly assigned level of education is impossible to measure. But that may not be a problem after all. As *Willis* and *Rosen* (1979) pointed out, this measure would have 'no significance as guides to the social or private profitability of investment in schooling' (*Willis–Rosen*, 1979, p. 11). The gains of a thorough education in econometrics with all necessary prerequisites would probably exceed its costs for most people, as they would find it a meaningless torture. On the other hand, people with appropriate interest, talent, and endurance probably find it very useful.

Instead of the mean return over all possible schooling levels for all people (the 'average treatment effect'), it makes sense to focus on the effect on those who have actually selected those (the 'average effect of the treatment on the treated'). The economic model of how people chose their education level helps identifying the problem.

A very simple model of education choice is enough to see why self-selection matters. In this model, individuals freely choose their schooling level. The only thing they care about is the present value of their lifetime earnings. They live an infinitely long life and face to a constant interest rate. They do not necessary face the same interest rate but it is constant throughout their lifetime. There are no costs of getting education except that they do not earn while in school. Schooling makes people earn more because it increases their marginal product. On the other hand, the increase in their marginal product is smaller and smaller as their schooling level rises. Again, we allow for heterogeneity in the relationship between marginal product and schooling.

The individuals' schooling choice is, therefore, the solution of an investment problem, where the value of the forgone earnings in the near future are weighted against the value of the increased earnings in the more distant future. The role of the discount rate is crucial. Let s_i denote the years spent in school by person i, r_i the individual-specific interest rate, and $y_i(s)$ the earnings function, also varying from individual to individual. We assume that $y'_i(s)>0$ and $y''_i(s)<0$.

$$s_i^* = \arg\max_s \int_{t=s}^{\infty} e^{-r_i t} y_i(s) dt = \arg\max_s \left(\frac{e^{-r_i s} y_i(s)}{r_i} \right).$$

The solution to this problem is given by the first-order condition (the second-order condition is satisfied by the concavity of y)

$$r_i = \left.\frac{y'_i(s)}{y_i(s)}\right|_{s=s_i^*} = \left.\frac{d\ln y_i(s)}{ds}\right|_{s=s_i^*}.$$

The 'returns to schooling' is defined as the value of the equality as we stated the first-order condition. It is the derivative of the logarithm of the earnings function at the optimum, which is equal to the discount rate of the individual faces. Since both the interest rate and the value of the derivative of the log earnings function can vary across people, the optimal schooling choice is expected to be different. When other things are equal, a higher r_i implies a lower optimal level of education. If two people differ solely in the interest rate they face, then they will have different schooling levels with the same earnings function $y(s)$.

In this case

$$\frac{\ln y_i(s_i^*) - \ln y_j(s_j^*)}{s_i^* - s_j^*} = \frac{\ln y_i(s_i^*) - \ln y_i(s_j^*)}{s_i^* - s_j^*} \approx \left.\frac{d\ln y_i(s)}{ds}\right|_{s=s^*},$$

so a simple reduced-form model (in this case a *Wald* regression) consistently estimates the causal effect, which is the same for both individuals. The estimator is consistent for the average treatment effect, that is the causal effect of randomly assigned education levels on people's earnings. The assignment is not random but that is not a problem, since it is uncorrelated with the effect (because the effect is the same for everybody). There may be a problem if the effect varies at different levels of education, but effects that are local to the different levels can be captured anyway. Obviously, the assumption of homogeneous earnings capacity is not realistic. If we allow for heterogeneous earnings functions the result does not hold anymore.

If two people face the same interest rate but have different earnings capacity, the one with a higher optimal level of schooling must have a higher $(\ln y_i)'$ at the optimum of the other:

$$s_i^* > s_j^* \quad \& \quad r_i = r_j \quad \Rightarrow \quad (\ln y_i)'\big|_{s=s_i^*} = (\ln y_j)'\big|_{s=s_j^*} \quad \Rightarrow \quad (\ln y_i)'\big|_{s=s_j^*} > (\ln y_j)'\big|_{s=s_j^*},$$

because $(\ln y_i)'$ is a decreasing function. For the same reason, we have that

$$\ln y_i(s_j^*) < \ln y_j(s_j^*),$$

and so

$$\frac{\ln y_i(s_i^*) - \ln y_j(s_j^*)}{s_i^* - s_j^*} > \frac{\ln y_i(s_i^*) - \ln y_i(s_j^*)}{s_i^* - s_j^*} \approx \left.\frac{d\ln y_i(s)}{ds}\right|_{s=s^*}.$$

The reduced-form model is biased upward. This is sometimes called the 'ability bias' of the reduced-form estimators, where heterogeneity in ability means heterogeneity of the earnings functions. Uncorrelated heterogeneity in interest rates has no effect on the bias, but there are special cases when a negative correlation might have an opposite effect (see the general setup by *Card*, 1999). In general, however, reduced form estimates will overstate the causal effect of education on earnings.

The model has two important implications. First, given people are free to chose the education they want given r_i and y_i, the derivative of the earnings function, that is, the causal effect of education on earnings can be observed only at the optimal level of education. Second, differences in earnings that correspond to different education levels overstate the causal effect of education.

The first point can be illustrated in the thought experiment to measure the causal effect of education on the earnings of a particular individual. This effect is the derivative of the $\ln y_i$ function, but one can only aim at measuring this derivative at the optimal s^*. The following is one appropriate design: let the individual choose an education level. Observe that: in according to the model, it is s^*. Also observe the corresponding earnings, $\ln y_i(s_i^*)$. Then induce a slightly different schooling level to the individual, and observe that s_i and the corresponding $\ln y_i(s_i)$. The individual has to change her decision, so a new s_i is going to be optimal. One way of doing this would be to change the interest rate she faces, another to constrain her choice to a slightly different previously non-optimal s_i. The difference between the two measured points would approximate $d\ln y_i(s_i^*)/ds$. This thought experiment identifies the local average effect of the treatment on the treated. It is not equal to the average treatment effect (the mean over all different education levels across all people) because for each individual, it is measured at the person-specific optimum.

In real life, of course, one person chooses some education level only once and gets lifetime earnings also once in a lifetime. Therefore, the only way to measure any effect is through inter-personal differences in schooling and earnings. The second implication of the model means that reduced form estimators like OLS overstate the causal effect, even in the local sense.

In addition, real-life measurement might contain measurement errors. Measurement error in the left-hand side variable does not affect consistency of the estimator, but right-hand side errors do. For the reason mentioned previously, measurement error in the schooling level variable has received considerable attention in the literature, and therefore it will be incorporated in the analysis. In the next section, the ways how these two econometric problems affect the OLS estimator and how a valid IV can help will be presented.

2. THE EMPIRICAL MODEL

Let s_i be the observed time (years) spent in school, an imperfect measure of real time spent in school, s_i^*. (The notation is a bit unfortunate: from now on, s_i is the observed value of the optimal choice s_i^*, not just any schooling level as before.) Let x_i be a k–1 dimensional vector of other factors affecting earnings, z_i a vector of factors affecting the schooling outcome but not the earnings, and y_i the logarithm of (lifetime) earnings.

With the measurement error, the model is specified by three equations:

$$y_i = \beta s_i^* + \delta' x_i + u_i \,, \qquad /1/$$

$$s_i^* = \alpha' z_i + \gamma' x_i + v_i \,, \qquad /2/$$

$$s_i = s_i^* + w_i \,, \qquad /3/$$

x_i and z_i are assumed to be uncorrelated with each of the error terms, u_i, v_i, and w_i.

Let us examine the inconsistency of the OLS estimator. OLS estimation of /1/ is inconsistent because s_i^* is endogenous and may be badly measured. s_i^* is endogenous because of a nonzero covariance of u_i and v_i, the unobserved heterogeneity in schooling assignment and earnings. A positive correlation is implied by the former simple model: u_i represents unobserved heterogeneity of the earnings functions, while v_i represents unobserved heterogeneity in the interest rate and the earnings function.

$$E\left[s_i^* u_i\right] = E\left[\left(\alpha' z_i + \gamma' x_i + v_i\right) u_i\right] = E\left[v_i u_i\right] = \sigma_{uv} \neq 0 \,. \qquad /4/$$

The measured model is:

$$y_i = \beta s_i - \beta w_i + \delta' x_i + u_i \equiv \beta s_i + \delta' x_i + \varepsilon_i \,, \qquad /5/$$

$$s_i = \alpha' z_i + \gamma' x_i + v_i + w_i = \alpha' z_i + \gamma' x_i + \eta_i \,. \qquad /6/$$

The measurement error, w_i is independent of all exogenous variables and all other error terms, by assumption. Therefore, z_i and x_i are uncorrelated with η_i and ε_i. Estimates of the earnings equation by OLS are inconsistent because of the covariance between the schooling level and the unobserved heterogeneity, and because of the measurement error. The asymptotic bias is a function of σ_{uv}, σ_w^2, and the moments of the covariates in the earnings equation (s and x). Let us derive the probability limit of the OLS estimator.

$$\begin{pmatrix} \hat{\beta} \\ \hat{\delta} \end{pmatrix}_{OLS} \equiv \left[\sum_{i=1}^{n} \begin{pmatrix} s_i^2 & s_i x'_i \\ x_i s_i & x_i x'_i \end{pmatrix} \right]^{-1} \sum_{i=1}^{n} \begin{pmatrix} s_i y_i \\ x_i y_i \end{pmatrix} = \left(\mathbf{X}'_n \mathbf{X}_n\right)^{-1} \mathbf{X}'_n \mathbf{y}_n =$$

$$= \left(\mathbf{X}_n^{*\prime} \mathbf{X}_n^* + \Sigma_w\right)^{-1} \left(\mathbf{X}_n^{*\prime} \mathbf{X}_n^* \begin{pmatrix} \beta \\ \delta \end{pmatrix} + \mathbf{X}_n^{*\prime} \mathbf{u}_n \right) \qquad /7/$$

where

$$\mathbf{X}_n \equiv \left[s_i \quad x_i' \right]_n \,, \qquad \mathbf{X}_n^* \equiv \left[s_i^* \quad x_i' \right]_n \,, \qquad \Sigma_n \equiv \begin{bmatrix} \sigma_w^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_n \,,$$

$$\mathbf{y}_n \equiv \left[y_i \right]_n \,, \qquad \mathbf{u}_n \equiv \left[u_i \right]_n \,.$$

Therefore,

$$plim\begin{pmatrix}\hat{\beta}\\ \hat{\delta}\end{pmatrix}_{OLS} = plim\left(\mathbf{X}_n^{*\prime}\mathbf{X}_n^{*} + \sigma_w \mathbf{e}_n \sigma_e \mathbf{e}_n'\right)^{-1}\left(\mathbf{X}_n^{*\prime}\mathbf{X}_n^{*}\begin{pmatrix}\beta\\ \delta\end{pmatrix} + \mathbf{X}_n^{*}\mathbf{u}_n\right) =$$

$$= \left(\Phi^{-1} - \frac{1}{1+\sigma_w^2\xi}\Phi^{-1}\Sigma_w\Phi^{-1}\right)\left(\Phi\begin{pmatrix}\beta\\ \delta\end{pmatrix} + E\left[\begin{pmatrix}s\\ x\end{pmatrix}u\right]\right) =$$

$$= \Phi^{-1}\left(\mathbf{I} - \frac{1}{1+\sigma_w^2\xi}\Sigma_w\Phi^{-1}\right)\left(\Phi\begin{pmatrix}\beta\\ \delta\end{pmatrix} + \begin{pmatrix}\sigma_{uv}\\ 0\end{pmatrix}\right),$$

where

$$\Phi \equiv E\left[\left(s^* x'\right)'\left(s^* x'\right)\right], \qquad \mathbf{e}_n = \left[\hat{\varepsilon}_i\right]_n$$

and $\xi>0$ is the upper left element of Φ^{-1}.

One can show that the previous expression implies that

$$\hat{\beta}_{OLS} \to \frac{\beta + \sigma_{uv}}{1+\sigma_w^2\xi} \qquad \text{in probability as } n \to \infty,$$

where

$$\xi \equiv E\left[s_i^{*2}\right] - E\left[s_i^* x_i'\right]E\left[x_i x_i\right]^{-1}E\left[x_i s_i^*\right].$$

The OLS estimator is asymptotically biased by the covariance of the structural error terms in an additive way: the sign of the bias is the same as the sign of σ_{uv}. The effect of the measurement error is different: it makes β biased toward zero. In the presence of a positive correlation ($\sigma_{uv}>0$) and measurement error, the two effects are of opposite direction if $\beta>0$.

3. THE INSTRUMENTAL VARIABLE (IV) STRATEGY

The most commonly used empirical strategy to address these econometric problems is instrumental variables estimation. This involves using one or more instruments that affect the schooling decision but are uncorrelated with earnings, conditional on education. In the thought experiment introduced before, good instruments are those that would induce some individuals to choose a different education level than they would choose otherwise. Obviously, the thought experiment itself cannot be carried out, but one can hope to find two groups of otherwise comparable individuals: one that was affected by the instrument, and another one that was not. The IV strategy gets around the endogeneity problem. Moreover, it is consistent regardless of potential measurement error in education.

In general, the IV strategy estimates the parameter of interest for the subpopulation which was affected by the instruments in the sense that they would have changed their

behaviour (*Imbens–Angrist*; 1994). The results are therefore local and they correspond to the effect of the treatment on the treated.

The validity of the instruments is an assumption: it cannot be inferred from the sample. Therefore, good instruments are not only hard to find but their quality is not testable. It is a matter of theoretical and speculative discussion.

Before turning to the alternative IV estimators, let us understand why weak instruments can be a problem. To keep things very simple, assume that z is one dimensional and there is no x vector. That is, we have one instrument, z, and one badly measured and endogenous variable, s^*:

$$y_i = \beta s_i^* + u_i \qquad /8/$$

$$s_i^* = \alpha z_i + v_i \qquad /9/$$

$$s_i = s_i^* + w_i \qquad /10/$$

Formally, the IV assumptions state that the instrument is uncorrelated with the structural error terms and have some effect on the schooling choice:

$$E[z_i u_i] = E[z_i v_i] = E[z_i w_i] = 0\,, \qquad E[z_i s_i^*] \neq 0\,. \qquad /11/$$

The IV estimator of β is

$$\hat{\beta}_{IV} \equiv \frac{\hat{\sigma}_{zy}}{\hat{\sigma}_{zs}} \qquad /12/$$

It is consistent for β:

$$plim\hat{\beta}_{IV} = \frac{E[zy]}{E[zs]} = \frac{\beta E[zs^*] + E[zu]}{E[zs^*] + E[zw]} = \beta\,.$$

Why a weak instrument is problematic is not difficult to see. We say that an instrument is weak if it does not predict well the endogenous variable, or, in other words, if the regression of s on z (the first stage regression) has an R^2 close to zero. For given σ_s^2 and σ_z^2, this means that σ_{zs} is small in the limit, and therefore so will be typically in finite samples. Since zero population covariances do not mean zero covariances in finite samples, we can have the IV estimator dominated by covariances with the structural errors if $E[z_i s_i^*]$ is small. The problem of weak instruments is therefore a finite-sample problem. By definition, finite-sample problems are not relevant if the sample is large enough. How large is large enough however depends on the particular problem. *Angrist and Krueger* (1991) use quarter and state of birth as an instrument of schooling level, arguing that state-specific compulsory schooling laws might affect final schooling levels. That is, people born in the fall start school almost a year later than those born in the summer, and therefore some of them might complete one less class than children of the summer. Their

sample has 329 000 observations but their instruments predict very little of actual schooling level (the partial R^2 is around 0.0001). Simulations by *Bound et al*. (1995) demonstrate that *Angrist* and *Krueger's* (1991) result can easily be an artifact of small-sample bias and tell nothing about the causal relationship. In fact, the causal relationship may well be zero. The simulation results are theoretically supported in the alternative asymptotics of *Staiger* and *Stock* (1997). The *F*-statistic on the excluded instruments in the first-stage regression is an indicator of how strong the instruments are. An *F*-statistic below 10 is usually seen as a sign of warning.

While intuitively appealing, this rule of thumb is not fully justified by econometric theory. A more careful way to detect weak instruments is by simulation exercises. They have an additional advantage in that they may tell something about the possible remedies if needed. I will present such an exercise on Harmon and Walker's estimator.

The IV estimators

If there are more than one excluded instrumental variables (z), there are more ways to combine them. In what follows, two different methods will be introduced. Three additional estimators will be defined as possible remedies for the weak instrument problem. They are all consistent (under the IV assumptions), but their asymptotic variance and their finite-sample bias and variance can be quite different.

Some new notation will help in the definitions. As before, bold letters denote sample matrices (and vectors). Let S denote the sample sum of squares, and θ the vector of parameters of interest, β and δ:

$$S_{ZZ} \equiv \frac{1}{n}\mathbf{Z}_n'\mathbf{Z}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} z_i' & x_i' \end{pmatrix}'\begin{pmatrix} z_i' & x_i' \end{pmatrix},$$

$$S_{ZX} \equiv \frac{1}{n}\mathbf{Z}_n'\mathbf{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} z_i' & x_i' \end{pmatrix}'\begin{pmatrix} s_i & x_i' \end{pmatrix},$$

$$S_{Zy} \equiv \frac{1}{n}\mathbf{Z}_n'\mathbf{y}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} z_i' & x_i' \end{pmatrix}' y_i ,$$

$$\theta \equiv \begin{pmatrix} \beta & \delta' \end{pmatrix}'.$$

Also, let $\mathbf{P}$ denote the projection matrix onto the column space of $\mathbf{Z}$, and $\mathbf{M}$ matrix creating the residual:

$$\mathbf{P}_n \equiv \mathbf{Z}_n\left(\mathbf{Z}_n'\mathbf{Z}_n\right)^{-1}\mathbf{Z}_n',$$

$$\mathbf{M}_n \equiv \mathbf{I}_n - \mathbf{P}_n .$$

The Optimal GMM Estimator

Generalized Method of Moments (GMM) estimators are based on moment restrictions.

Here those involve the covariance of the excluded instruments and the error term in the earnings equation:

$$E[z_i'\varepsilon_i] = E[z_i'(y_i - \beta s_i - \delta' x_i)] = 0$$

The GMM in general is defined in the following way:

$$\hat{\theta}_{GMM} \equiv \left(S'_{ZX}\, A^{-1} S_{ZX}\right)^{-1} S'_{ZX}\, A^{-1} S_{Zy} \qquad /13/$$

The estimator is consistent with any positive definite matrix A. The optimal GMM estimator is a special case, where A, the weighting matrix, is the covariance matrix of the product of the error term and the instruments (in the broad sense):

$$\hat{\theta}_{OGMM} \equiv \left(S'_{ZX}\, \Omega^{-1} S_{ZX}\right)^{-1} S'_{ZX}\, \Omega^{-1} S_{Zy}\,, \qquad /14/$$

where

$$\Omega \equiv Var\left[\varepsilon \begin{pmatrix} z \\ x \end{pmatrix}\right] = E\left[\varepsilon^2 \begin{pmatrix} z \\ x \end{pmatrix} \begin{pmatrix} z' & x' \end{pmatrix}\right].$$

The implementation requires an estimate of Ω. Since the estimator is consistent with any other positive definite matrix, in the first step one can get a consistent estimate for ε_i by using any appropriate matrix. In particular, the identity matrix meets the required condition. The optimal GMM estimation therefore consists of two steps. The first one consists of estimating $\hat{\theta}_{OGMM_1} \equiv \left(S'_{ZX}\, S_{ZX}\right)^{-1} S'_{ZX}\, S_{Zy}$, and taking the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_{OGMM_1}$. Ω can then be estimated using the estimated residuals and **Z**. The second step is the optimal GMM estimation, by using $\hat{\Omega}$. The optimal GMM estimator is the minimum distance combination of the different instruments if the system is overidentified. It is optimal in the sense that it has the smallest asymptotic variance.

Two Stage Least Squares

The Two Stage Least Squares (2SLS) estimator of θ is defined by

$$\hat{\theta}_{2SLS} \equiv \left(S'_{ZX}\, S_{ZZ}{}^{-1} S_{ZX}\right)^{-1} S'_{ZX}\, S_{ZZ}{}^{-1} S_{Zy}\,. \qquad /15/$$

This estimator is equivalent to the two-stage procedure of first regressing s on all z and x ('first stage'), and then regressing y on the first-stage prediction of s and the x variables ('second stage').

$\hat{\theta}_{2SLS}$ is a GMM estimator with weight matrix S_{ZZ}^{-1}, and therefore it is consistent. Moreover, it is equivalent to the optimal GMM estimator if ε is homoscedastic. With

out homoscedasticity, however, the two estimators are going to be different, and therefore 2SLS is not optimal. In cross-sectional applications, homoscedasticity is a rare exception. On the other hand, the 2SLS estimator is a lot simpler to compute, and it is part of all major statistical packages. For this reason, it remains the most popular IV estimator.

The k-class estimators

k-class estimators are a generalization of the two-stage least square introduced by *Theil* (1958), and are defined as:

$$\hat{\theta}_k \equiv \left[\mathbf{X}_n'\mathbf{P}_n\mathbf{X}_n - (1-k)\mathbf{X}_n'\mathbf{M}_n\mathbf{X}_n\right]^{-1}\left[\mathbf{X}_n'\mathbf{P}_n\mathbf{y}_n - (1-k)\mathbf{X}_n'\mathbf{M}_n\mathbf{y}_n\right]. \qquad /16/$$

2SLS is a *k*-class estimator, with *k*=1. *Nagar*'s estimator (introduced by *Nagar*; 1959) is another special case, where

$$k_{Nagar} = 1 + \frac{q-2}{n} \qquad /17/$$

q being the dimension of *z*, that is the number of instruments excluded from the earnings equation. Nagar's estimator has the minimum expected bias in finite samples among the *k*-class estimators. One of the assumptions behind this result is that the x_i are nonstochastic. *Donald* and *Newey* (1997) generalize the Nagar results and suggest a modified version

$$k_{Donald-Newey} = 1 + \frac{q-2}{n}\Big/\left(1 - \frac{q-2}{n}\right). \qquad /18/$$

Limited Information Maximum Likelihood

Some more notation. Let **Y** be the sample matrix of the endogenous variables, and **P**(*x*) be the projection onto $[x_i']_n$:

$$\mathbf{Y}_n \equiv [y_i \quad s_i]_n, \quad \mathbf{P}(x) \equiv [x_i']_n\left([x_i']_n'[x_i']_n\right)^{-1}[x_i']_n', \quad \mathbf{M}(x) \equiv \mathbf{I}_n - \mathbf{P}(x).$$

The Limited Information Maximum Likelihood (LIML) estimator is derived from the likelihood function assuming normal errors. It can be expressed, however, as another *k*-class estimator with $k=\lambda$, where λ is the smallest eigenvalue of the matrix B defined as

$$\mathbf{B} \equiv \left(\mathbf{Y}_n'\mathbf{M}_n\mathbf{Y}_n\right)^{-1}\mathbf{Y}_n'\mathbf{M}(x)\mathbf{Y}_n.$$

Staiger and *Stock* (1997) show that in their framework, the LIML estimator has the best finite-sample properties.

4. HARMON AND WALKER'S STUDY

Harmon and *Walker* (1995) follow an instrumental variables approach to estimate β on British data. They use a pooled sample of the consecutive cross-sectional waves of the British Family Expenditures Survey (1978–1986). The sample is quite large, $n = 34\,336$. They use the cohort of the individual as an instrument. Their motivation is the following. The minimum school-leaving age was increased two times in the relevant period from 14 to 15 in 1947, and from 15 to 16 in 1973. They provide evidence that these changes indeed changed the behaviour of many individuals: quite a few of those who otherwise would have left school stayed on because of the new law. This strategy is an example of what the natural experiments literature calls a difference in differences estimator (*Meyer*; 1995).

The instrument directly changes the schooling level 'assignment' of the individuals who complete one more class during this time, since it forces them to choose $s_i > s_i^*$, similarly to our thought experiment. It measures the treatment on the treated, those that would have left school at a younger age but had had to stay in school and completed one more class. The instrument is strong enough if this sub-population is significant.

Technically, their instrument is a 3-valued vector indicating whether the person is a member of the cohort that was subject to the first, second, or third minimum school-leaving age (SLA) requirement (SLA=14, 15, or 16). z_i therefore is a vector 2 binary variables, taking SLA=14 the reference group. The other covariates, the x_i-s include a constant, age, age squared, region (10 categories + 1 reference), and year of survey (8 categories + 1 reference). An important implication of the 2-dimensional z_i vector is that 2SLS is equivalent to Nagar's estimator ($k_{Nagar}=1+(q-2)/n$) and also the *Donald–Newey* estimator. Therefore, in this empirical model, the 2SLS estimator is the IV estimator with the best finite-sample properties in the *Nagar* (1959) and in the *Donald* and *Newey* (1997) setup. It is dominated by the LIML estimator in the framework of *Staiger* and *Stock* (1997).

The OLS estimate is a lot smaller than the IV:

$$\hat{\beta}_{OLS} = 0.06\,,\ \hat{\beta}_{2SLS} = 0.15\,.$$

Taken by face value, the estimated causal effect of education is very large: it states that an additional year spent in school increases earnings by 15 percent for people at the very bottom end of the schooling distribution. This is a surprisingly large effect (most other estimates are at most 10 percent), and one has to be sure that the results are not confounded by small-sample effects or other factors before taking it seriously.

Provided that the IV estimate is correct, one can also conclude that OLS is biased downward by 60 percent. The question our simulation will answer is what combination of endogeneity and measurement error is needed for this result.

The Monte Carlo exercise

To examine the finite-sample properties of the Harmon and Walker's 2SLS estimator and the alternative IV estimators, this paper generated data similar to the original sample. The artificial samples were drawn from a population with moments that match the re

ported ones. The Data Generating Process (DGP) of the part of z_i (i.e. the two binary SLA variables) is a multinomial process. The DGP of x_i consists of age (a uniform random variable in the simulations), its square, and two sets of multinomial variables for the mutually exclusive categories of the region and the year of the survey. The covariance of x with z was preserved (except for some simplifying assumptions with negligible consequences). The appendix table compares the simulated moments to the published ones.

The DGP for the structural error terms (u_i and v_i) was modeled as bivariate normal with a correlation of ρ_{uv}. s_i^* was generated following the schooling equation with the reported parameters. Observed schooling attainment variable was $s_i = s_i^* + w_i$, $w_i \sim iid\ N(0,\sigma_w^2)$, representing the measurement error. Finally, y_i was generated by the earnings equation, again, with the published parameters. Several combinations of ρ_{uv} and σ_w^2 were examined.

Formally, the generated vectors z_i and x_i and scalars u_i, v_i, and w_i were used to generate y_i, s_i^*, and s_i the following way:

$$y_i = \beta s_i^* + \delta' x_i + u_i ,$$
$$s_i^* = \alpha' z_i + \gamma' x_i + v_i ,$$
$$s_i = s_i^* + w_i ,$$

where Harmon and Walker's point estimates were used for β, δ, α, and γ. In each run of the simulation new data were generated, and $\hat{\beta}$ was estimated in each dataset. The purpose of the Monte Carlo experiment was to examine the properties of four different estimators (OLS, optimal GMM, 2SLS and LIML) by comparing then to the 'theoretical' value of beta used in the data generating process.

Multiple measures of the bias were considered: the mean error (Bias) and the mean squared error (MSE) are reported in the following, the median and mean absolute error are available upon request. The reported measures are defined as

$$Bias \equiv \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\beta}_j - \beta\right), \qquad MSE \equiv \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\beta}_j - \beta\right)^2 ,$$

where M is the number of Monte Carlo replications, and $\hat{\beta}_j$ is the estimate of β in the j-th replication.

The following tables summarize the results of twelve simulations, each with different parametrization of the econometric problems ρ_{uv} and σ_w^2, each from 1000 replications. With no econometric problems (ρ_{uv} =0, σ_w^2=0), OLS is the best. IV estimators beat OLS in terms of the bias even with relatively small problems (ρ_{uv} =0.05, or σ_w^2=0.1). The difference among the alternative IV estimators are small. All of these indicate that Harmon–Walker's estimator delivers the asymptotic results. LIML seems to slightly outperform the 2SLS (which is here equivalent to the Nagar and also the Donald–Newey estimator), but this result is not robust.

Besides these results, I also carried out simulations on smaller samples. These might be interesting because of two reasons. First, it helps to see what sample size is 'large

enough' and in what sense it is large in our setup. Second, the relative performance of the 2SLS (Nagar) and the LIML estimator would probably show better results in smaller samples. The results for $n = 1\,500$, $n = 3\,000$, $n = 6\,000$, $n = 9\,000$, and $n = 15\,000$, are available on request. The simulations suggest that the IV estimators dominate the OLS estimator in terms of bias even with weak endogeneity or measurement error, and even in relatively small samples ($n = 1\,500$). Their superiority disappears in terms of MSE. In fact, a rather serious econometric problem is needed to get smaller MSE for any of the IV estimators, if the sample size is small: $\rho_{uv} = 0.2$ if $n = 1\,500$, and $\rho_{uv} = 0.1$ if $n = 3\,000$. Note that with any reasonable scale of measurement error, all IV estimators underperform OLS in an MSE sense, unless the sample size is very large ($n > 10\,000$).

The small-sample results are not clear about the relative performance of the 2SLS (Nagar and Donald–Newey) and the LIML estimators in terms of the bias. On the other hand, the MSE and mean absolute error results seem rather robust. The LIML estimator underperforms 2SLS (Nagar) in small samples in terms of these measures. These results are consistent with the general notion that, even if it is better in expectation, the variance of the LIML estimator does not converge to zero (the asymptotic variance of the root-n magnified estimator is infinite).

Comparison of the performance of different estimators

Endogeneity (ρ_{uv})	Measurement Error (σ_w^2)	OLS	OGMM	2SLS	LIML
		Bias			
0.00	0.00	0.000	0.000	0.000	0.000
0.05	0.00	0.049	-0.001	-0.001	-0.001
0.10	0.00	0.098	0.000	0.000	0.000
0.15	0.00	0.147	0.000	0.000	0.000
0.20	0.00	0.195	-0.001	-0.001	-0.001
0.00	0.10	-0.025	0.000	0.000	0.000
0.00	0.20	-0.046	-0.001	-0.001	-0.001
0.00	0.50	-0.099	-0.001	-0.001	-0.001
0.10	0.10	0.057	-0.003	-0.003	-0.003
0.20	0.10	0.014	0.002	0.002	0.002
-0.10	0.10	-0.107	-0.002	-0.001	-0.001
-0.20	0.10	-0.190	-0.002	-0.001	-0.001
		MSE			
0.00	0.00	0.0000	0.0017	0.0017	0.0017
0.05	0.00	0.0024	0.0013	0.0013	0.0013
0.10	0.00	0.0096	0.0016	0.0016	0.0016
0.15	0.00	0.0216	0.0016	0.0016	0.0016
0.20	0.00	0.0382	0.0012	0.0012	0.0012
0.00	0.10	0.0006	0.0017	0.0017	0.0017
0.00	0.20	0.0022	0.0013	0.0013	0.0013
0.00	0.50	0.0097	0.0015	0.0015	0.0015
0.10	0.10	0.0033	0.0014	0.0014	0.0014
0.20	0.10	0.0195	0.0015	0.0015	0.0015
-0.10	0.10	0.0115	0.0014	0.0014	0.0014
-0.20	0.10	0.0360	0.0015	0.0015	0.0015

5. CONCLUSION

Harmon and Walker's IV estimate is more than twice larger than their OLS result not because their instruments are weak. Their model may give misleading results, though, for a different reason. The instrument they use is people's birth cohort. As it is often the case with similar difference in different natural experiment estimators, there are other things that might be varied between those groups, other than the compulsory schooling laws they faced. For one thing, the quadratic age-earnings profile may not be the right specification to use even if differences in age purely reflect variety in labor market experience (see *Murphy–Welch*; 1990). Moreover, in this cross-sectional setup, the joint effect of age and birth cohort on earnings may reflect time effects in earnings (cycles and trends). Another possible problem is the time path of employment in the unskilled group they focus on: their employment rate fell through the time considered in the developed world, therefore probably in Britain, too. That might have introduced selection problems into observed wages because the sample became more 'able'.

If we accept the results these problems notwithstanding, they tell us that those that would have left school at the minimum school age (14 or 15) but stayed one year longer because of the new law earn 15 per cent more because of this extra year. This is a classical local effect of the treatment on the treated. Generalizing these effects to the rest of the population is not justified by the empirical model itself: we need some theory for that. However, in the model presented previously (or, for that matter, in *Card* 1999), this is not possible without making more structure on the heterogeneity in the costs of education and the earnings function. The large literature on educational choice has not provided enough evidence for that yet.

The instrument in the illustrative example stood well in the simulation exercise but may be problematic for other reasons. I believe that one should carry out a similar analysis if there is enough reason to worry about the finite sample properties of the estimators, and one should check the robustness of the estimators to the other problems. Moreover, one has to be explicit in what exactly the results mean in terms of the locality of the causal effect and the characteristics of the treatment group. There is no free lunch in econometrics either: the real benefits of the instrumental variables estimation strategy can be exploited only by careful analysis.

APPENDIX

Sample moments from Harmon and Walker (1995) and the simulated DGP (1000 replications)

Variable	Whole Sample		SLA=14		SLA=15		SLA=16	
	H&W	Simul	H&W	Simul	H&W	Simul	H&W	Simul
Ln(wage)	1.913	1.907	1.902	1.837	1.995	2.003	1.584	1.644
Std(lnwage)	0.445	1.035	0.434	3.296	0.416	1.983	0.426	4.126
Age	38.7	38.7	55.8	55.8	35.6	35.6	21.6	21.6
Std(age)	12.7	12.8	4.5	6.9	7.3	3.0	2.7	5.2
Region1	0.088	0.088	0.088	0.088	0.085	0.088	0.101	0.088
Region2	0.110	0.110	0.105	0.110	0.109	0.110	0.119	0.110

(Continued on the next page.)

(Continuation.)

Variable	Whole Sample		SLA=14		SLA=15		SLA=16	
	H&W	Simul	H&W	Simul	H&W	Simul	H&W	Simul
Region3	0.075	0.075	0.073	0.075	0.074	0.075	0.082	0.075
Region4	0.099	0.099	0.103	0.099	0.099	0.099	0.090	0.099
Region5	0.037	0.037	0.038	0.037	0.037	0.037	0.032	0.037
Region6	0.306	0.306	0.317	0.306	0.300	0.306	0.311	0.306
Region7	0.074	0.074	0.070	0.074	0.074	0.074	0.080	0.074
Region8	0.051	0.051	0.050	0.051	0.050	0.051	0.054	0.051
Region9	0.089	0.089	0.081	0.089	0.101	0.089	0.050	0.089
Region10	0.013	0.013	0.013	0.013	0.013	0.013	0.017	0.013
Year1	0.116	0.116	0.143	0.143	0.118	0.118	0.062	0.062
Year2	0.116	0.116	0.140	0.140	0.116	0.116	0.075	0.075
Year3	0.121	0.121	0.129	0.129	0.122	0.122	0.102	0.102
Year4	0.118	0.118	0.113	0.113	0.119	0.119	0.118	0.118
Year5	0.101	0.101	0.084	0.084	0.105	0.105	0.113	0.113
Year6	0.104	0.104	0.091	0.091	0.103	0.103	0.135	0.135
Year7	0.101	0.101	0.069	0.069	0.102	0.102	0.159	0.159
Year8	0.102	0.102	0.058	0.058	0.100	0.100	0.192	0.192

REFERENCES

ANGRIST, J. D. – KRUEGER, A. B. (1991): Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, No.106. p. 979–1014.

ANGRIST, J. D.– IMBENS, G. W. – RUBIN, D. B (1996): Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, Vol. 91, No. 434. p.444–455.

BOUND, J. – JAEGER, D. A. – BAKER, R. M. (1995): Problems with instrumental variables estimation when the correlation between the instrument and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, Vol. 90, No.430. p. 443–450.

CARD, D. (1999): The causal effects of education on earnings. In: ASHENFELTER, O. – CARD, D. (eds): *Handbook of labor economics*, Vol. 3A. Elsevier, North-Holland, p. 1801–1868.

DONALD, S. G. – NEWEY, K. W. (1997): *Choosing the number of instruments*. MIT Department of Economics. Working Paper.

HECKMAN, J. J. (1997): Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources,* Vol. 32. No. 3. p. 441–462.

IMBENS, G. W. – ANGRIST, J. D. (1994): Identification and estimation of local average treatment effects. *Econometrica,* Vol. 62. No.4. p. 467–476.

HARMON, C. – WALKER, I. (1995): Estimates of the economic return to schooling for the United Kingdom. *American Economic Review*, Vol. 85. No. 5. p. 1278–1286.

MEYER, B. (1995): Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, Vol. 13. No. 2. p. 151–161.

MURPHY, K. M. – WELCH, F. (1990): Empirical age-earnings profiles. *Journal of Labor Economics*, Vol. 8. No. 2. p. 202–229.

STAIGER, D. – STOCK, J. H. (1997): Instrumental variables regressions with weak instruments. *Econometrica*, Vol. 65. No. 3. p. 557–586.

THEIL, H. (1958): *Economic forecasts and policy*. North-Holland, Chapter 6.

WILLIS, R. J. – ROSEN, S. (1979): Education and self-selection. *The Journal of Political Economy*, Vol. 87. No. 2. p. S7–S36.

WILLIS, R. J. (1986): Wage determinants: A survey and interpretation of human capital earnings functions. In: ASHENFELTER, O. – LAYARD (eds): *Handbook of labor economics*, Vol. 1. p. 525–602.