Review of Economic Studies (1997) 64, 555–574 0034-6527/97/00260555$02.00


# Estimating Outcome Distributions for Compliers in Instrumental Variables Models

GUIDO W. IMBENS
*Harvard University and Arizona State University*

and

DONALD B. RUBIN
*Harvard University*

*First version received January* 1994; *final version accepted November* 1996 (*Eds.*)

In Imbens and Ingrist (1994), Angrist, Imbens and Rubin (1996) and Imbens and Rubin (1997), assumptions have been outlined under which instrumental variables estimands can be given a causal interpretation as a local average treatment effect without requiring functional form or constant treatment effect assumptions. We extend these results by showing that under these assumptions one can estimate more from the data than the average causal effect for the subpopulation of compliers; one can, in principle, estimate the entire marginal distribution of the outcome under different treatments for this subpopulation. These distributions might be useful for a policy maker who wishes to take into account not only differences in average of earnings when contemplating the merits of one job training programme vs. another. We also show that the standard instrumental variables estimator implicitly estimates these underlying outcome distributions without imposing the required nonnegativity on these implicit density estimates, and that imposing nonnegativity can substantially alter the estimates of the local average treatment effect. We illustrate these points by presenting an analysis of the returns to a high school education using quarter of birth as an instrument. We show that the standard instrumental variables estimates implicitly estimate the outcome distributions to be negative over a substantial range, and that the estimates of the local average treatment effect change considerably when we impose nonnegativity in any of a variety of ways.

## 1. INTRODUCTION

In recent empirical work (e.g. Angrist (1990), Angrist and Krueger (1991), Kane and Rouse (1992), Butcher and Case (1993), Card (1993), McClellan and Newhouse (1994)) researchers have attempted to estimate causal effects using instrumental variables to deal with possible self-selection into a treatment. Although there is a long tradition in cross-section econometrics of using instrumental variables estimation in self-selection problems (e.g. Gronau (1974), Willis and Rosen (1979), Heckman and Robb (1985)), this recent work, part of the *natural experiments* literature, differs from the older instrumental variables literature in its increased focus on the validity of the instruments, often at the expense of the strength of the relation between the instrument and the endogenous regressor. In Imbens and Angrist (1994), Angrist and Imbens (1995), Angrist, Imbens and Rubin (1996) and Imbens and Rubin (1997), assumptions have been outlined under which such instrumental variables estimands can be given a causal interpretation as a *local average treatment effect* without requiring functional form or constant treatment effect assumptions.

In this paper we make two points: first we show that under these assumptions one can estimate more from the data than the average causal effect for the subpopulation of *compliers*; one can, in principle, estimate the entire marginal distribution of the outcome under the different treatments for this subpopulation. These distributions might be useful for a policy maker who wishes to take into account not only differences in average earnings but also differences in dispersion of earnings when contemplating the merits of one programme or treatment vs. another.

Second, we show that the standard instrumental variables estimator implicitly estimates these underlying outcome distributions without imposing the required nonnegativity condition on density estimates, and that imposing nonnegativity on these implicit density estimates can substantially alter the estimates of the local average treatment effect.

We illustrate these points in two ways. First we present an analysis of the returns to a high school education using quarter of birth as an instrument. We show that the standard instrumental variables estimates implicitly estimate the outcome distributions to be negative over a substantial range, and that the estimates of the local average treatment effect change considerably when we impose nonnegativity in any of a variety of ways. Second, we do a small Monte Carlo study to show that the proposed estimators that impose nonnegativity on the outcome distributions can have substantially lower root mean squared error than the standard IV estimator.

## 2. IDENTIFICATION OF THE LOCAL AVERAGE TREATMENT EFFECT USING INSTRUMENTAL VARIABLES

In this section we set up the framework used to analyse instrumental variables estimators. The "potential outcome" framework we use is based on Rubin's (1974, 1978, 1990) extension of Neyman's (1923) model for randomized experiments to observational settings including possible interference between units and versions of each treatment, and allowing outcomes to be stochastic, specialized to instrumental variables in Angrist, Imbens and Rubin (1996). Following Holland (1986) we refer to this as the Rubin Causal Model (RCM).

Let $Z_i$ be a binary instrument. Let the pair $D_i(0)$ and $D_i(1)$ denote the values of the treatment for individual $i$ that would be obtained given the instrument $Z_i=0$ and $Z_i=1$ respectively. If $D_i(0)=0$ and $D_i(1)=1$ unit $i$ is called a *complier*. For $z=0, 1$ and $d=0, 1$, let $Y_i(z, d)$ denote the outcome that would be observed given instrument $Z_i=z$ and treatment $D_i=d$ respectively. Implicit in this notation is the Stable Unit Treatment Value Assumption (SUTVA, Rubin (1980, 1990)), which requires that unit $i$ is not affected by the treatment received and instrument assigned for other units. We also make the standard econometric instrument or exclusion assumption that the potential outcomes $Y_i(z, d)$ do not depend on $z$; for any unit there are, therefore, only two different potential outcomes $Y_i(d)$, one for each value of the treatment $D_i$: $Y_i(0)$ is the outcome that would be observed if the treatment were $D_i=0$, and $Y_i(1)$ is the outcome that would be observed if the treatment were $D_i=1$. The third assumption is the strict monotonicity assumption which requires that $D_i(1)\geqq D_i(0)$ for all units $i$, with inequality for at least one unit. This assumption requires that changing the instrument from $Z_i=0$ to $Z_i=1$ would not lead anyone to shift from receiving the treatment to not receiving the treatment; that is, there are no units with $D_i(0)=1$ and $D_i(1)=0$. (Labelled *defiers* by Balke and Pearl (1993).) Finally, we assume that the instrument $Z_i$ is randomly assigned, independent of the potential outcomes $D_i(0)$, $D_i(1)$, $Y_i(0)$ and $Y_i(1)$, or more generally, ignorable (Rubin (1978)). For all individuals we observe the triple $Z_{\text{obs},i}=Z_i$, $D_{\text{obs},i}=D_i(Z_{\text{obs},i})$ and $Y_{\text{obs},i}=Y_i(D_{\text{obs},i})$.

Under these assumptions (SUTVA, the exclusion restriction, strict monotonicity and random assignment of the instrument) one can estimate the local average treatment effect, the average of the unit level treatment effect, $Y_i(1) - Y_i(0)$, for the subpopulation of compliers characterized by $D_i(0) = 0$ and $D_i(1) = 1$, by taking the ratio of the average difference in $Y_i$ by instrument and the average difference in $D_i$ by instrument

$$E[Y_i(1) - Y_i(0) | D_i(0) = 0, D_i(1) = 1] = \frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]},$$

where $E[\cdot]$ denotes population averages. In Angrist, Imbens and Rubin (1996) these assumptions are discussed in detail and some examples in which these assumptions may be justified are given.

A more conventional econometric approach starts with the switching regression model (Maddala and Nelson (1975), Willis and Rosen (1979), Bjorklund and Moffitt (1987), Heckman (1990)), where two outcomes are postulated:

$$Y_i(0) = \beta_0 + \varepsilon_i, \tag{1}$$

$$Y_i(1) = \gamma_0 + \xi_i, \tag{2}$$

in combination with a latent variable describing the selection:

$$D_{\text{obs},i} = 1\{\pi_0 + \pi_1 Z_{\text{obs},i} + \nu_i > 0\}, \tag{3}$$

where $1\{\cdot\}$ is the indicator function, equal to one if its argument is true and zero otherwise, and with the observed outcome equal to $Y_{\text{obs},i} = Y_i(0) \cdot (1 - D_{\text{obs},i}) + Y_i(1) \cdot D_{\text{obs},i}$. The key assumption is that $Z_i$ is independent of all disturbances $\varepsilon_i$, $\xi_i$ and $\nu_i$. These models are typically estimated under additional distributional assumptions using maximum likelihood methods because instrumental variables estimation is not consistent for the average treatment effect $\gamma_0 - \beta_0$ that is typically the focus in program evaluation.[1]

A special case of the switching regression model is the dummy endogenous variable-constant coefficient model characterized by the equation

$$Y_{\text{obs},i} = \beta_0 + \beta_1 D_{\text{obs},i} + \varepsilon_i, \tag{4}$$

combined with equation (3).[2] In this constant treatment effect model instrumental variables is consistent for the treatment effect $\beta_1$.

In another version of the dummy endogenous variable model, the participation equation is not explicitly written down. Instead, the response equation is presented together with the two assumptions that (i), $\varepsilon_i$ is uncorrelated with $Z_i$ and (ii), there is a non-zero correlation between $D_i$ and $Z_i$. Although the set-up is weaker than assuming full independence of $\varepsilon_i$ and $Z_i$, with possible dependence between $\varepsilon_i$ and $Z_i$ a zero correlation implies that a variable $Z_i$ could be a valid instrument for the effect of $D_i$ on $Y_i$, without being a valid instrument for the effect of $D_i$ on a transformation of $Y_i$ such as $\ln Y_i$. Because part of the appeal of the natural experiment literature is in its lack of reliance on functional form assumptions, we do not regard this as an appealing relaxation of the assumptions.

1. An exception is Heckman (1990) who presents identification results requiring the support of the instrument $Z_i$ to be unbounded.

2. See for example Heckman (1978).

## 3. IDENTIFICATION OF THE MARGINAL OUTCOME DISTRIBUTIONS FOR COMPLIERS

A policy maker or individual decision maker may be interested in more than just average treatment effects. For example, a policy maker contemplating a training programme may be interested in the proportion of the population whose earnings will be above the poverty level given the training relative to the proportion with above poverty level earnings given no training. Alternatively, the policy maker may be interested in the effect of the training programme on earnings equality as measured by the variance of earnings. In all these cases, knowledge of the distribution of earnings given training and the distribution of earnings given no training would enable the policy maker to answer these questions. In this section we show that given the four assumptions discussed in Section 2, SUTVA, the exclusion restriction, strict monotonicity and randomization of the instrument, one can estimate for compliers the distribution of outcomes both given treatment and given no treatment. In order to focus on this identification issue, the discussion in this section assumes that the joint distribution of observables $(Z_{\text{obs},i}, D_{\text{obs},i}, Y_{\text{obs},i})$ is estimated without sampling error.

Although we can identify the two marginal outcome distributions for compliers, we cannot, under our assumptions, identify the joint distribution of $Y_i(0)$ and $Y_i(1)$ for compliers or the distribution of their individual gains $Y_i(1) - Y_i(0)$. This is, of course, not possible in a randomized experiment either, and it can be argued in that context that in many cases the two marginal distributions comprise all that is of interest.[3] To pursue this point briefly, consider an individual contemplating taking one of two treatments. In this decision process it may be of use to evaluate the distribution of outcomes for "comparable" individuals under both treatments. Specifically, suppose that in a randomized experiment 50% of the individuals exposed to treatment A improved and 60% of the individuals exposed to treatment B improved. One can think of four types of individuals, depending on whether they would improve or not given treatment A and improve or not given treatment B. Knowledge of the joint distribution of outcomes amounts to knowing both the two marginal outcome distributions as well as the population distribution of the four types. One distribution of types that is consistent with the results of the randomized experiment is that 10% of the experimental population improve under treatment B but not under treatment A, and nobody improves under treatment A but not under treatment B. A second distribution of types consistent with the evidence is that 40% of the experimental population improve under A but not under B and 50% improve under B but not under A. Unless the individual decision maker has at least partial knowledge about which of the four types she is, in which case she should only consider the experiences of similar type individuals and disregard experiences of different types, there appears to be no relevance to the decision maker of knowing the type distribution in the population. It can therefore be argued that within subpopulations of units that are exchangeable with respect to observable characteristics, there is no useful information (in the sense of affecting decisions) in the joint outcome distribution that is not contained in the two marginal distributions. Information about the distribution of variables that are not observed cannot be used for conditioning in decision making and therefore can often be regarded as superfluous.

In cases where we are interested in individual outcomes the correlation between the two potential outcomes may be of interest. Consider the case of a person who has been exposed to a treatment, say a drug, and for whom we have observed an outcome, say

3. See Clements, Heckman and Smith (1994) for a different view.

death. It might be of interest, for example in a legal setting, to consider whether the person would have died had he not been exposed to the treatment. Answering this question would require knowledge about the joint distribution that cannot fully be gleaned from the two marginal distributions.

To discuss the identification of the marginal outcome distributions for compliers it is convenient to introduce additional notation. We partition the population by the effect of the treatment assignment on treatment received; for *never-takers* (units with $D_i(0)=0$, $D_i(1)=0$), let $C_i=n$; for *always-takers* (units with $D_i(0)=1$ and $D_i(1)=1$), let $C_i=a$; finally for *compliers* (units with $D_i(0)=0$ and $D_i(1)=1$), let $C_i=c$. These three types exhaustively partition the population since by the monotonicity assumption there are no *defiers* (units with $D_i(0)=1$ and $D_i(1)=0$).[4]

Let $\phi_n$, $\phi_a$ and $\phi_c$ denote the population frequencies of the three types of individuals. Only compliers (units with $C_i=c$) are affected by the instrument; the local average treatment effect is the average causal effect for this subpopulation. We cannot directly learn anything about the causal effects of $D$ on $Y$ for always-takers because we never observe them without the treatment, and we cannot directly learn anything about the causal effect of $D$ on $Y$ for never-takers because we never observe them with the treatment. Although we might well be interested in causal effects for these groups, any estimates of average causal effects for them, and therefore any estimates of the population average causal effects, require additional information or assumptions about their responses to treatments to which they are never observed to be exposed.[5]

If we were to observe the population type, inference would be straightforward: ignoring all noncompliers we would compare outcomes in the two treatment groups for the subpopulation of compliers. However, because an individual's type is not always identifiable from the observed variables, inference must be indirect, based on treatment groups that are mixtures of compliers and non-compliers.

Although we cannot generally identify the compliers from the observed data ($Z_{\text{obs},i}$, $D_{\text{obs},i}$, $Y_{\text{obs},i}$), we can identify some of the non-compliers; if $Z_{\text{obs},i}=0$ and $D_{\text{obs},i}=1$, then individual $i$ must be an always-taker with $C_i=a$, and if $Z_{\text{obs},i}=1$, and $D_{\text{obs},i}=0$, then individual $i$ must be a never-taker with $C_i=n$. Because of randomization, the instrument $Z_{\text{obs},i}$ is independent of $(D_i(0), D_i(1))$ and therefore of $C_i$. Hence, in large samples we know the distribution of $Y_i(1)$ for always-takers; this distribution will be denoted by $g_a(y)$. Analogously, in large samples, we know the distribution of $Y_i(0)$ for never-takers; this distribution will be denoted by $g_n(y)$. Note that because we assume full independence rather than mean independence of the instrument $Z_i$ and the potential outcomes $Y_i(0)$ and $Y_i(1)$, these distributions are not indexed by the value of the instrument. By the independence of instrument $Z_i$ and type $C_i$, in large samples we also know the population proportions of the types: $\phi_n=\Pr(D_{\text{obs},i}=0|Z_{\text{obs},i}=1)$, $\phi_a=\Pr(D_{\text{obs},i}=1|Z_{\text{obs},i}=0)$ and thus we can deduce $\phi_c=1-\phi_n-\phi_a$.

4. In terms of the selection equation (3) these three types can be defined as

$$C_i=\begin{cases} n & \text{if} \quad v_i \leqq -\pi_0-\pi_1 \\ c & \text{if} \quad -\pi_0-\pi_1 < v_i \leqq -\pi_0 \\ a & \text{if} \quad -\pi_0 < v_i. \end{cases}$$

This illustrates that the monotonicity assumption which asserts that there are no defiers with $D_i(1)=0$ and $D_i(0)=1$, is made implicitly rather than explicitly in the equation-based model despite its critical importance for causal inference.

5. This information can be in the form of bounds on the range of outcomes. See Robins (1989), Manski (1990), and Balke and Pearl (1993) for calculations of bounds on treatment effects in the presence of such information.

It remains to find the two critical outcome distributions, the distributions of $Y_i(0)$ and $Y_i(1)$ among compliers; call these $g_{c0}(y)$ and $g_{c1}(y)$. These are more complicated to find from observed data than the distributions for the non-compliers because among those assigned $Z_{\text{obs},i}=0$, both never-takers and compliers will be observed to have $D_{\text{obs},i}=0$. Analogously, in the subsample with $Z_{\text{obs},i}=1$, compliers and always-takers will be observed to have $D_{\text{obs},i}=1$.

At this point some additional notation is useful. Let $f_{zd}(y)$ denote the directly estimable distribution of $Y_{\text{obs},i}$ in the subsample defined by $Z_{\text{obs},i}=z$ and $D_{\text{obs},i}=d$. We will write the distributions of interest in terms of these directly estimable distributions. As already noted, $g_n(y)=f_{10}(y)$ and $g_a(y)=f_{01}(y)$. Individuals assigned to $Z_{\text{obs},i}=0$ and exposed to $D_{\text{obs},i}=0$ are a mixture of compliers and never-takers. By independence of instrument $Z_i$ and type $C_i$, the sampling distribution $f_{00}(y)$ is a mixture of the distribution of $Y_i(0)$ for never-takers, $g_n(y)$, and for compliers, $g_{c0}(y)$, with the mixing probability equal to the relative probability of these subpopulations in the entire population

$$f_{00}(y)=\frac{\phi_c}{\phi_c+\phi_n}g_{c0}(y)+\frac{\phi_n}{\phi_c+\phi_n}g_n(y).$$

Analogously, for individuals assigned to $Z_{\text{obs},i}=1$ and exposed to $D_{\text{obs},i}=1$, we can rule out that such individuals are never-takers, but we cannot infer whether these individuals are always-takers or compliers. The distribution of the outcome in this subsample is therefore a mixture of the population distribution of $Y_i(1)$ for compliers, $g_{c1}(y)$, and for always-takers, $g_a(y)$, with the mixing probability equal to the relative population proportions of these two subpopulations

$$f_{11}(y)=\frac{\phi_c}{\phi_c+\phi_a}g_{c1}(y)+\frac{\phi_a}{\phi_c+\phi_a}g_a(y).$$

The four directly estimable distributions $f_{zd}(y)$ have now been expressed in terms of the two complier distributions of interest, $g_{c0}(y)$ and $g_{c1}(y)$, and the two directly estimable nuisance distributions $g_n(y)=f_{10}(y)$ and $g_a(y)=f_{01}(y)$ for never-takers and always-takers, respectively. We can invert these relations and express the two potential outcome distributions of interest in terms of the observable distributions

$$g_{c0}(y)=\frac{\phi_n+\phi_c}{\phi_c}f_{00}(y)-\frac{\phi_n}{\phi_c}f_{10}(y), \tag{5}$$

and,

$$g_{c1}(y)=\frac{\phi_a+\phi_c}{\phi_c}f_{11}(y)-\frac{\phi_a}{\phi_c}f_{01}(y). \tag{6}$$

Thus, from the four directly estimable distributions, we can derive the entire complier distribution of outcomes under each value of the treatment, $g_{c0}(\cdot)$ and $g_{c1}(\cdot)$, rather than just the difference in their means, which is the instrumental variables estimand.

## 4 THE ANATOMY OF CONVENTIONAL INSTRUMENTAL VARIABLES ESTIMATES

In this section we show that standard instrumental variables estimates are implicitly based on estimates of the two complier outcome distributions that are not restricted to be nonnegative. We then show that this point can have important implications for inference because restricting these estimates to be nonnegative, even in a naive way, can change inference considerably, as we illustrate in an example where we estimate earnings returns to high school using quarter of birth as an instrument.

To further investigate the conventional instrumental variables estimator, define $\bar{Y}_{zd}$ to be the average of the observed outcome $Y_{\text{obs},i}$ for the subsample with observed instrument $Z_{\text{obs},i}=z$ and observed treatment $D_{\text{obs},i}=d$

$$\bar{Y}_{zd}=\sum_{i=1}^{N} Y_{\text{obs},i}1\{Z_{\text{obs},i}=z\}1\{D_{\text{obs},i}=d\}/\sum_{i=1}^{N} 1\{Z_{\text{obs},i}=z\}1\{D_{\text{obs},i}=d\}.$$

In addition define

$$\bar{D}_z=\sum_{i=1}^{N} D_{\text{obs},i}1\{Z_{\text{obs},i}=z\}/\sum_{i=1}^{N} 1\{Z_{\text{obs},i}=z\},$$
$$\bar{Y}_{\cdot d}=\sum_{i=1}^{N} Y_{\text{obs},i}1\{D_{\text{obs},i}=d\}/\sum_{i=1}^{N} 1\{D_{\text{obs},i}=d\},$$

and

$$\bar{Y}_{z\cdot}=\sum_{i=1}^{N} Y_{\text{obs},i}1\{Z_{\text{obs},i}=z\}/\sum_{i=1}^{N} 1\{Z_{\text{obs},i}=z\}=\bar{Y}_{z1}\bar{D}_z+\bar{Y}_{z0}(1-\bar{D}_z).$$

Then we can write the conventional IV estimator as

$$\hat{\beta}_{\text{IV}}=\frac{\bar{Y}_{1\cdot}-\bar{Y}_{0\cdot}}{\bar{D}_1-\bar{D}_0}=\hat{Y}_{c1}-\hat{Y}_{c0}, \tag{7}$$

where $\hat{Y}_{c1}=(\bar{D}_1\bar{Y}_{11}-\bar{D}_0\bar{Y}_{01})/(\bar{D}_1-\bar{D}_0)$ and $\hat{Y}_{c0}=((1-\bar{D}_0)\bar{Y}_{00}-(1-\bar{D}_1)\bar{Y}_{10})/(\bar{D}_1-\bar{D}_0)$. To interpret $\hat{Y}_{c1}$ and $\hat{Y}_{c0}$, consider the probability limits of the components of expression (7), $\bar{Y}_{00}$, $\bar{Y}_{01}$, $\bar{Y}_{10}$, $\bar{Y}_{11}$, $\bar{D}_0$, and $\bar{D}_1$. As argued before, the fraction of individuals with $D_{\text{obs},i}=1$ in the subsample with $Z_{\text{obs},i}=0$, $\bar{D}_0$, estimates the population share of the always-takers, $\phi_a$. Similarly, the fraction of individuals with $D_{\text{obs},i}=1$ in the subsample with $Z_{\text{obs},i}=1$, $\bar{D}_1$, estimates the combined population share of the always-takers and compliers, $\phi_a+\phi_c$. The denominator in (7), $\bar{D}_1-\bar{D}_0$, is therefore an unbiased estimate of $\phi_c$.

For each $(z, d)$, the expectation of $\bar{Y}_{zd}$ is equal to $E[Y_{\text{obs},i}|Z_{\text{obs},i}=z, D_{\text{obs},i}=d]$. We can use the relation between the directly estimable distributions $f(\cdot)$ and the distributions of interest $g(\cdot)$ to express these expectations of observed $Y_{\text{obs},i}$ conditional on observed instrument $Z_{\text{obs},i}$ and observed treatment $D_{\text{obs},i}$ in terms of the expectations of the potential outcomes $Y_i(1)$ and $Y_i(0)$ conditional on type $C_i$

$$\begin{aligned}
E[\bar{Y}_{00}]&=E[Y_{\text{obs},i}|Z_{\text{obs},i}=0, D_{\text{obs},i}=0]\\
&=\frac{\phi_c}{\phi_c+\phi_n}E[Y_i(0)|C_i=c]+\frac{\phi_n}{\phi_c+\phi_n}E[Y_i(0)|C_i=n],\\
E[\bar{Y}_{01}]&=E[Y_{\text{obs},i}|Z_{\text{obs},i}=0, D_{\text{obs},i}=1]=E[Y_i(1)|C_i=a],\\
E[\bar{Y}_{10}]&=E[Y_{\text{obs},i}|Z_{\text{obs},i}=1, D_{\text{obs},i}=0]=E[Y_i(0)|C_i=n],\\
E[\bar{Y}_{11}]&=E[Y_{\text{obs},i}|Z_{\text{obs},i}=1, D_{\text{obs},i}=1]\\
&=\frac{\phi_c}{\phi_c+\phi_a}E[Y_i(1)|C_i=c]+\frac{\phi_a}{\phi_c+\phi_a}E[Y_i(1)|C_i=a].
\end{aligned}$$

Inverting these relations we have

$$E[Y_i(0)|C_i=c]=\frac{E[\bar{D}_1]E[\bar{Y}_{11}]-E[\bar{D}_0]E[\bar{Y}_{01}]}{E[\bar{D}_1]-E[\bar{D}_0]}, \tag{8}$$

and

$$E[Y_i(1)|C_i=c]=\frac{E[1-\bar{D}_0]E[\bar{Y}_{00}]-E[1-\bar{D}_1]E[\bar{Y}_{10}]}{E[\bar{D}_1]-E[\bar{D}_0]}. \tag{9}$$

The first term on the right-hand side of (7), $\hat{Y}_{c1}$, is therefore an estimate of the expectation of the potential outcome given the treatment for the compliers and the second term, $\hat{Y}_{c0}$, an estimate of the expectation of the potential outcome without treatment for the compliers.

The first point, exemplifying the discussion in Section 3, is that we can directly obtain estimates of $E[Y_i(1)|C_i=c]$ and $E[Y_i(0)|C_i=c]$ separately, not just of their difference $E[Y_i(1)-Y_i(0)|C_i=c]$. Such estimates might be informative about the plausibility of the assumptions underlying the instrumental variables estimates, and lead to a better understanding of the selection process, as will be demonstrated in the next section. The second point is that these estimates do not take into account the underlying mixture structure implied by the model. More precisely, these moment estimates do not take into account the fact that the two distributions $f_{00}(\cdot)$ and $f_{11}(\cdot)$ are mixtures of $g_n(\cdot)$ and $g_{c0}(\cdot)$ and of $g_a(\cdot)$ and $g_{c1}(\cdot)$, respectively, and densities must be nonnegative. In the next section we look at an empirical example and show that these restrictions are indeed important and can lead to substantially different inferences.

## 5. INSTRUMENTAL VARIABLES ESTIMATES OF THE RETURNS TO HIGH SCHOOL

As an illustration of the issues raised in this paper, we examine instrumental variables estimates of the returns to education. In an influential paper Angrist and Krueger (1991) (AK henceforth) investigated the causal effect of education on earnings. They noted that achieved education levels differed by date of birth for people born in a given year. They attributed this to compulsory schooling laws, which affect people born in different months of the same year in different ways: children start school at different ages but since they are all required to stay in school only until their sixteenth birthday, people are effectively faced with different amounts of compulsory schooling. AK then used the assigned amount of compulsory schooling as the instrument for achieved education levels. Since this is perfectly correlated (within each state) with season of birth, this strategy is referred to by AK as using "quarter of birth" as an instrument. AK discuss in detail why they believe this leads to a valid instrument.

We simplify the data AK investigated by making both treatment and instrument binary. The treatment of interest is defined as the indicator whether an individual has twelve or more years of education or less than twelve years of education, loosely corresponding to having a high school degree vs. not having a high school degree. This redefinition of the treatment raises an issue about the validity of the instrument. If quarter of birth is a valid instrument for years of education, it is not necessarily a valid instrument for a treatment defined as a function of years of education such as the indicator function we are using. This issue is also relevant for the AK study: if quarter of birth is a valid instrument for education measured in months, it is not necessarily a valid instrument for the level of education rounded off to the nearest year. Although the approach is still straightforward with a multivalued treatment, the number of types increases rapidly with the number of distinct levels of the treatment, leading to a mixture structure with the number of mixture components $g(\cdot)$ exceeding the number of directly estimable distributions $f(\cdot)$. Modelling decisions will necessarily be more important in that case.

TABLE 1

*Average outcome ($Y_i$ is log weekly earning) by instrument ($Z_i$ is indicator for born in fourth quarter) and treatment status ($D_i$ is indicator for twelve years of schooling)*

| | $D_{\text{obs},i}=0$ | $D_{\text{obs},i}=1$ | Total | $\bar{D}_i$ |
|---|---|---|---|---|
| $Z_{\text{obs},i}=0$ | $\bar{Y}_{00}=5{\cdot}595$ ($N=19{,}454$) | $\bar{Y}_{01}=5{\cdot}984$ ($N=62{,}217$) | $\bar{Y}_{0\cdot}=5{\cdot}892$ ($N=81{,}671$) | $\bar{D}_0=0{\cdot}762$ ($N=81{,}671$) |
| $Z_{\text{obs},i}=1$ | $\bar{Y}_{10}=5{\cdot}597$ ($N=17{,}632$) | $\bar{Y}_{11}=5{\cdot}991$ ($N=63{,}212$) | $\bar{Y}_{1\cdot}=5{\cdot}905$ ($N=80{,}844$) | $\bar{D}_1=0{\cdot}782$ ($N=80{,}844$) |
| Total | $\bar{Y}_{\cdot 0}=5{\cdot}596$ ($N=37{,}086$) | $\bar{Y}_{\cdot 1}=5{\cdot}988$ ($N=125{,}429$) | $\bar{Y}=5{\cdot}898$ ($N=162{,}515$) | $\bar{D}=0{\cdot}772$ ($N=162{,}515$) |

We only consider people born in the first or fourth quarters, thereby reducing the instrument to a binary one. The extension to multivalued instruments is straightforward: the monotonicity assumption is required to hold for any pair of values of the instrument thereby leading to a more complicated mixture structure with the number of types equal to the number of distinct values of the instrument plus one. All distributions of interest can be recovered from the joint distribution of the observed variables.

The data we use are taken from the AK study and comprise observations from the 1980 census on weekly earnings, years of education and quarter of birth for 162,515 white men born between 1 January 1930 and 31 December 1939 during the first or fourth quarter of each year. In Table 1 we present the averages and sample sizes for the different values of treatment and instrument ($D_{\text{obs},i}=0$ implies less than twelve years of education, $D_{\text{obs},i}=1$ implies at least twelve years of education. $Z_{\text{obs},i}=0$ implies birth in first quarter, and $Z_{\text{obs},i}=1$ implies birth in fourth quarter).

The treatment-control average difference, which is identical to the ordinary least squares estimate of the returns to high school, is $\bar{Y}_{\cdot 1}-\bar{Y}_{\cdot 0}=5{\cdot}988-5{\cdot}596=0{\cdot}391$. The conventional instrumental variables estimate, which is the ratio of differences of average outcome by instrument status to the difference in average treatment probability by instrument, is $(\bar{Y}_{1\cdot}-Y_{0\cdot})/(\bar{D}_1-\bar{D}_0)=(5{\cdot}905-5{\cdot}892)/(0{\cdot}782-0{\cdot}762)=0{\cdot}651$.

It is interesting to note that, similar to what has been found in a number of studies where returns to education have been estimated using instrumental variables (Angrist and Krueger (1991), Butcher and Case (1993), Card (1993)), we find that this instrumental variables estimate of the returns to education is considerably larger than the corresponding ordinary least squares estimate—the difference in averages by treatment status. In contrast, in the earlier literature on returns to education (see Griliches (1977) for a discussion), it was often hypothesized that ordinary least squares estimates of the return to education over-estimated the causal effect of education on earnings because of the so-called "ability bias". This bias was hypothesized to reflect a propensity of people with high ability and high earnings potential to have levels of education higher on average than those of people with low ability and low earnings potential. Card (1993) and others have pointed to these recent instrumental variables results as an indication that if anything, least squares estimates underestimate the returns to education.

To shed further light on this issue, we consider the additional information we can learn from the data about the outcome distributions for the compliers, $g_{c0}(y)$ and $g_{c1}(y)$. Their estimated means, based on the implicit estimates $\hat{Y}_{c0}$ and $\hat{Y}_{c1}$ in the standard instrumental variables approach, are 5·57 and 6·23 for $Y_i(0)$ and $Y_i(1)$ respectively. Comparing these to the estimated means of $Y_i(0)$ for never-takers (5.57) and of $Y_i$ (1) for always-takers (5·99), we see that the difference between the treatment-control difference (or ordinary least squares estimate) of 0·39 and instrumental variables estimate of 0·65 is entirely due to the

large difference between the estimated mean of $Y_i(1)$ for compliers and always-takers, $6{\cdot}23-5{\cdot}99=0{\cdot}24$, with the difference between the estimated mean of $Y_i(0)$ for compliers and never-takers equal to zero.

In Figures 1–4 we give histogram estimates of the distribution of observed outcome by instrument and treatment status, $\hat{f}_{zd}(y)$ for $z=0, 1$ and $d=0, 1$, with the binwidth fixed at 0·1. The differences between the directly estimable $\hat{f}_{00}(y)$ and $\hat{f}_{10}(y)$ and between $\hat{f}_{01}(y)$ and $\hat{f}_{11}(y)$ are barely noticeable. This reflects the fact that the instrument $Z_i$ is a very weak one, in the sense that $Z_i$ is only very weakly correlated with $D_i$, the treatment of interest: the estimate of the average causal effect of quarter of birth on receiving at least twelve years of education, which is an unbiased estimate of the population proportion of compliers $\phi_c$, is only two percent.

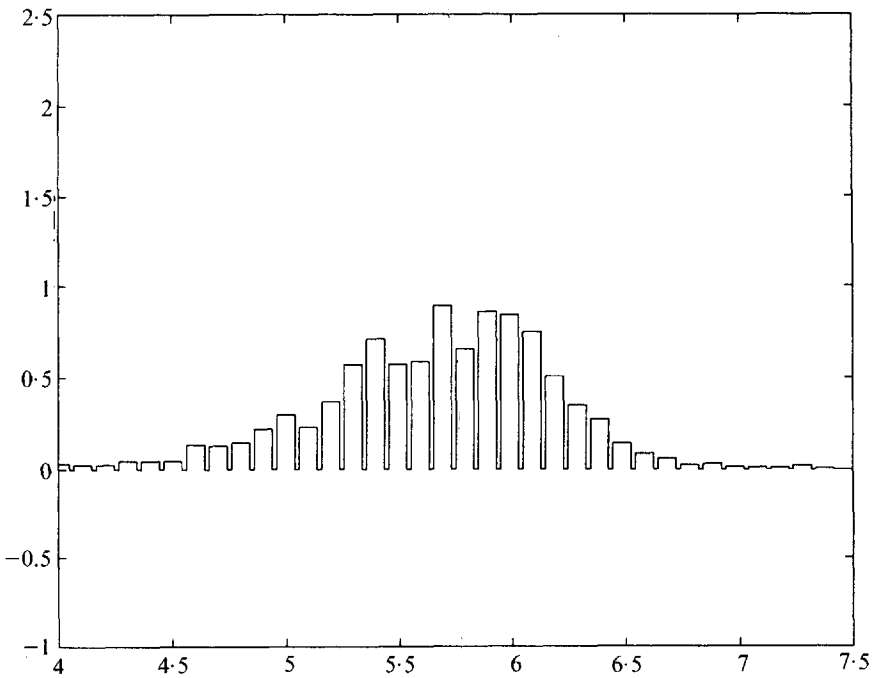


FIGURE 1
Histogram for $f_{00}$

In Figure 5–6 we present simple unrestricted histogram-type estimates of the two complier distributions $\hat{g}_{c0}(y)$, and $\hat{g}_{c1}(y)$ based on equations (5)–(6) using the histogram estimates for the four sampling distributions $\hat{f}(\cdot)$, and estimates for the proportions of the different types of $\hat{\phi}_c=0{\cdot}020$, $\hat{\phi}_n=0{\cdot}218$ and $\hat{\phi}_a=0{\cdot}762$ respectively

$$\hat{g}_{c0}(y)=\frac{\hat{\phi}_n+\hat{\phi}_c}{\hat{\phi}_c}\times\hat{f}_{00}(y)-\frac{\hat{\phi}_n}{\hat{\phi}_c}\times\hat{f}_{10}(y)=11{\cdot}90\times\hat{f}_{00}(y)-10{\cdot}90\times\hat{f}_{10}(y)$$

$$\hat{g}_{c1}(y)=\frac{\hat{\phi}_a+\hat{\phi}_c}{\hat{\phi}_c}\times\hat{f}_{11}(y)-\frac{\hat{\phi}_a}{\hat{\phi}_c}\times\hat{f}_{01}(y)=39{\cdot}10\times\hat{f}_{11}(y)-38{\cdot}10\times\hat{f}_{01}(y).$$

The bins for the histograms are $(v_{l-1}, v_l)$, for $l=1, \ldots, L$ where $v_0=3$, $v_l-v_{l-1}=0{\cdot}1$, $v_L=8{\cdot}5$, and $L=55$. The instrumental variables estimand is the difference in means of the

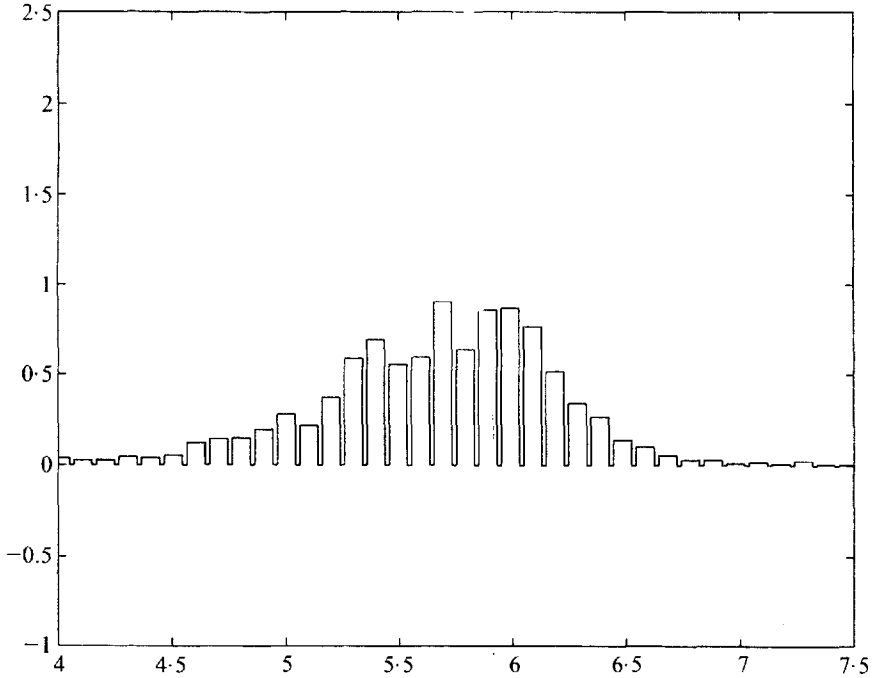


FIGURE 2
Histogram for $f_{10}$

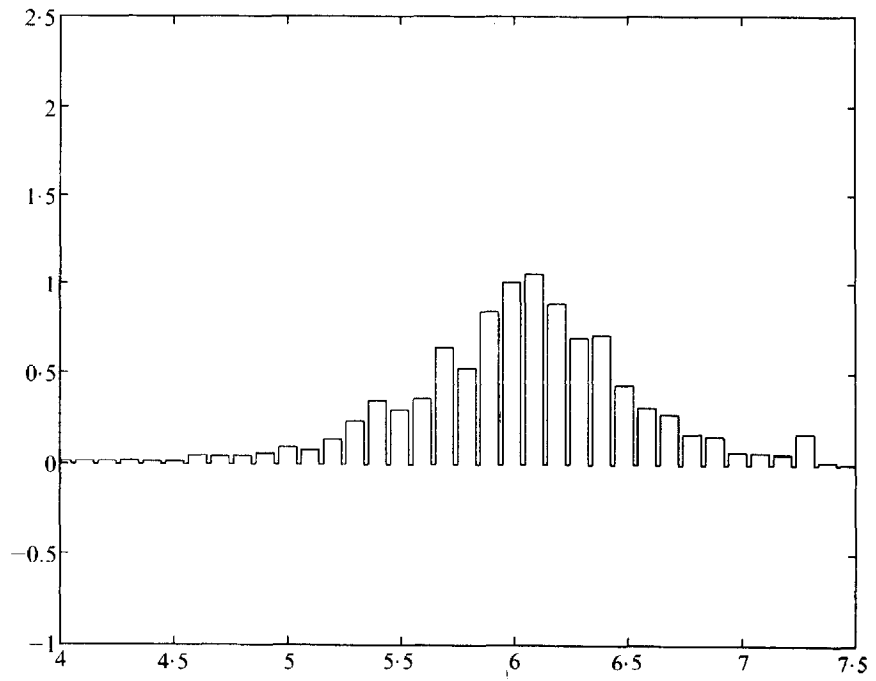


FIGURE 3
Histogram for $f_{01}$

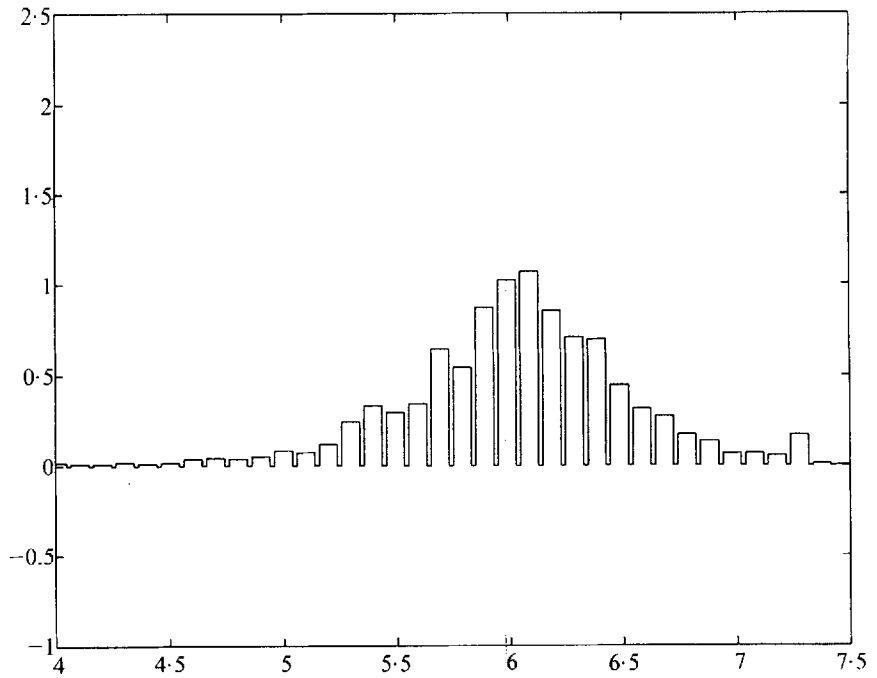


FIGURE 4
Histogram for $f_{11}$

two distributions $g_{c1}(y)$ and $g_{c0}(y)$ where the standard IV estimate is essentially a unrestricted estimate of the difference in the two means. If we integrate the two histogram estimates $\hat{g}_{c1}(y)$ and $\hat{g}_{c0}(y)$ of the density functions, we obtain 0·66 for the difference. The slight difference between this and the standard IV estimate of 0·65 is due to the smoothing implicit in the histogram estimates with non-negligible binwidth of the density functions.

These estimates of the entire distribution of the potential outcomes for the compliers allow the further interpretation of the difference between the ordinary least squares estimate, that is, the treatment-control difference, and the instrumental variables estimate. Inspecting the distributions, presented in Figures 5 and 6, that underlie these estimates reveals the fact that the last two histogram estimates of the two compliers distributions are not everywhere nonnegative. The estimate of the distribution of $Y_i(0)$ for compliers, $\hat{g}_{c0}(y)$, in Figure 5, does not suffer much from this, and the estimate is comparable to the estimate of the distribution of $Y_i(0)$ for never-takers, $g_n(y)=f_{10}(y)$, in Figure 2. In contrast, the estimate of the distribution of $Y_i(1)$ for compliers, $g_{c1}(y)$, in Figure 6, does seem quite different from the estimate of the distribution of $Y_i(1)$ for always-takers, $g_a(y)=f_{01}(y)$ in Figure 3, and is negative over a large range of values. This negativity can be due to sampling variation or to violations of the assumptions. In particular, it can point to violations of the exclusion assumption. If the exclusion assumption is violated, and there is a direct effect of the instrument on the outcome, there is no reason to expect this particular linear combination of the sampling distributions to be nonnegative. It is also possible that negativity points to violations of the monotonicity assumption or the randomization of $Z$, although these assumptions seem plausible in this context. See AK for a discussion of the plausability of the ignorability assumption in this application, and Angrist and Imbens (1995) for a discussion of the monotonicity condition.

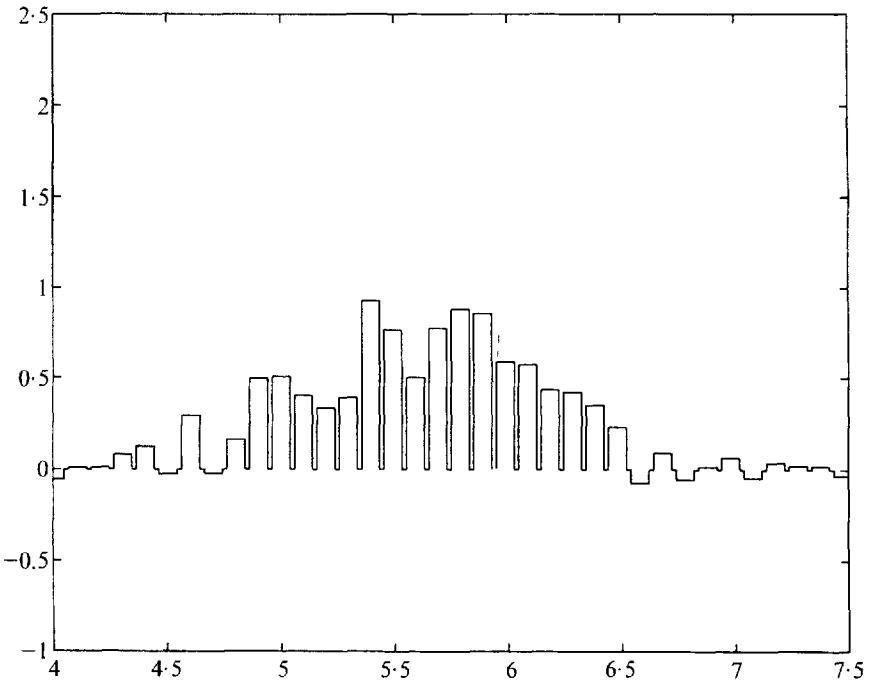


FIGURE 5
Histogram for $g_0^*$

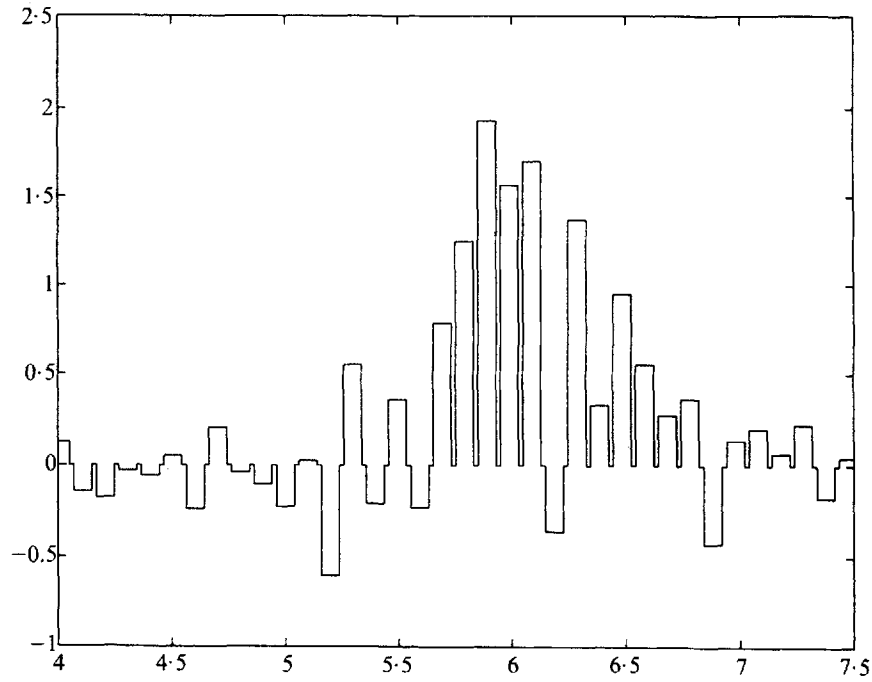


FIGURE 6
Histogram for $g_1^*$

## 6. ALTERNATIVE ESTIMATES OF THE RETURNS TO HIGH SCHOOL

In this section we present three new alternatives to the standard IV estimates of the returns to high school with quarter of birth as an instrument, each of which keeps distributional estimates in the proper parameter space, that is, ensure that the estimates of the two complier density functions $g_{c1}(y)$ and $g_{c0}(y)$ are nonnegative. The first two estimators are, in spirit, nonparametric estimators, where we model each of the four density functions in a flexible way as piece-wise constant with 55 pieces for a total of 282 parameters. The third estimator assumes normality for the four outcome distributions with a total of eight parameters. More generally we view this third estimator as an example of the type of parametric estimator one may wish to use in practice. Although the identification results in Section 3 ensure that in principle nonparametric estimation is possible, in small samples more parsimonously parametrized models based on the normal distribution or generalizations, e.g. the $t$-distribution or a mixture of normal distributions, may do a better job of smoothing the data and lead to a smaller root-mean-squared-error at the expense of some bias as we shall show in Section 7. The role of the parametric model is solely to provide a good fit to the four underlying outcome distributions. The estimand of interest, the difference in means of the two complier distributions, is well defined irrespective of the specific parametric model used.

The first estimator, which we refer to as the "nonnegative" IV estimator, is a slight modification of the histogram estimates discussed in the previous section. Let $\hat{g}_{c0}^{\text{pos}}(y)$ and $\hat{g}_{c1}^{\text{pos}}(y)$ denote the estimates

$$\hat{g}_{cj}^{\text{pos}}(y) = \max(0, \hat{g}_{cj}(y)) \left[ \int \max(0, \hat{g}_{cj}(y)) dy \right]^{-1},$$

for $j=0, 1$, where $\hat{g}_{c0}(y)$ and $\hat{g}_{c1}(y)$ are the implicit IV distribution estimates discussed in the previous section. The estimates for the two noncomplier distributions are the same as before: $\hat{g}_n^{\text{pos}}(y) = \hat{g}_n(y)$ and $\hat{g}_a^{\text{pos}}(y) = \hat{g}_a(y)$.

The second estimator is the maximum likelihood estimator based on four multinomial models with constant density on intervals $v_{l-1}$ to $v_l$ for $l=1, \ldots, L$, where $v_0 = 3$, $v_l - v_{l-1} = 0{\cdot}1$, $v_L = 8{\cdot}5$, and $L = 55$, thereby forcing the nonnegativity restrictions to be satisfied by choosing $\hat{g}_{cj}^{\text{ml}}(y)$, $\hat{g}_a^{\text{ml}}(y)$, and $\hat{g}_n^{\text{ml}}(y)$ to maximize the likelihood function rather than adjusting only the complier distributions as the nonnegative IV estimator does. Note that, in order to maintain comparability, the bins $(v_l, v_{l-1})$ are the same for both the nonnegative IV and the multinomial ML estimators. Within the framework of this discrete approximation to the four outcome distributions, the restrictions are inequality restrictions in a parametric model.

The third estimator is the maximum likelihood estimator with the four outcome distributions normal with unknown means and variances. We impose the restriction that the variance of $Y_i(0)$ for compliers equals that for never-takers and the variance of $Y_i(1)$ for compliers equals that for always-takers. Calculation of the maximum likelihood estimates is based on the EM algorithm (Dempster, Laird and Rubin (1977)).

Table 2 presents estimates of the mean and variance of the four outcome distributions as well as estimates of the average effect for compliers for the standard IV model (using the implicit estimates for $E[Y_i(0)|C_i=c]$ and $E[Y_i(1)|C_i=c]$ given in (7)) and for the three alternatives just introduced, with standard errors based on large-sample normal approximations. All three new alternatives lead to estimates of the local average treatment effect substantially smaller than the standard IV estimate.

TABLE 2

*Estimates of mean and variance of potential outcomes*

| | Never-takers $Y_i(0)$ | | Always-takers $Y_i(1)$ | | $Y_i(0)$ | | Compliers $Y_i(1)$ | | $Y_i(1)-Y_i(0)$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | var | mean | var | mean | var | mean | var | mean | s.e. |
| Standard IVE | 5·59 | 0·52 | 5·99 | 0·42 | 5·57 | 0·43 | 6·23 | 0·17 | 0·66 | (0·17) |
| Nonnegative IVE | 5·59 | 0·52 | 5·98 | 0·41 | 5·48 | 0·97 | 5·94 | 1·23 | 0·45 | (0·17) |
| Multinomial MLE | 5·60 | 0·49 | 5·99 | 0·40 | 5·49 | 0·97 | 5·92 | 1·23 | 0·42 | (0·17) |
| Normal MLE | 5·60 | 0·51 | 5·99 | 0·41 | 5·53 | 0·51 | 6·03 | 0·41 | 0·50 | (0·15) |

We can also compare the estimates of the entire density functions to those obtained for the standard IV estimator. In Figures 7 and 8 we present the ML estimates of the density functions of $g_{c0}(y)$ and $g_{c1}(y)$ respectively under the multinomial and normal models. The nonnegative IV estimates are essentially identical to the multinomial maximum likelihood estimates and therefore not separately displayed.

The estimates of the first two moments in Table 2 for the alternative procedures and the density estimates in Figures 7 and 8 tell a markedly different story from the conventional IV estimates. They suggest that the distribution of $Y_i(1)$ for compliers is not as different from the distribution of $Y_i(1)$ for always-takers as suggested by the standard IV estimates. For example, the mean for $Y(1)$ for compliers, implicitly estimated by the standard IV procedure as 6·23, is estimated by the other three procedures to be between 5·92 and 6·03, much closer to the estimate of the average of $Y(1)$ for always-takers (5·99). The compliers are very similar to the noncompliers with the same level of education. Although one many argue with the choice of the three alternative estimators, the fact that

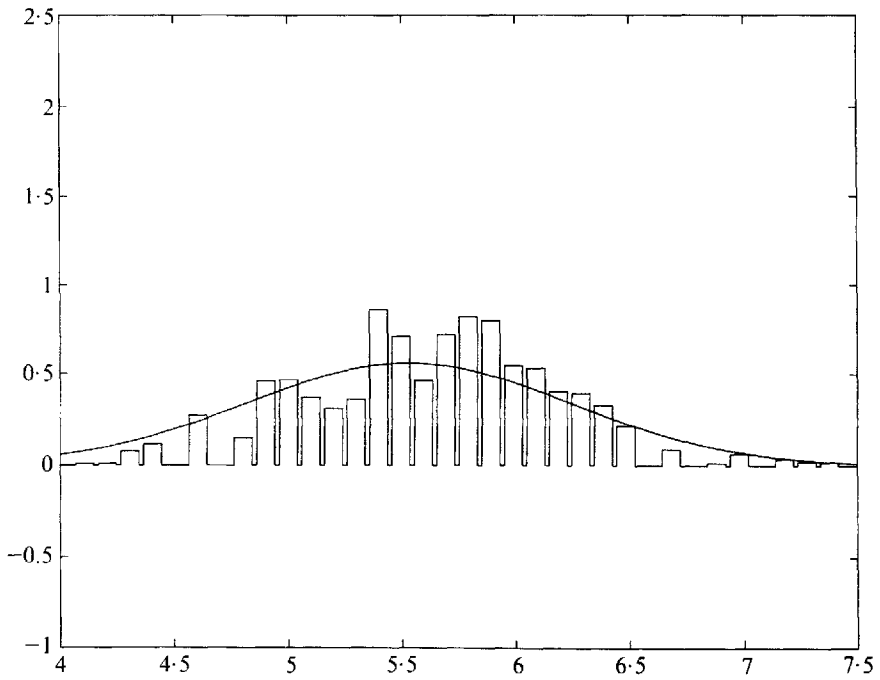


FIGURE 7

Multinomial and normal MLEs for $g_0^*$

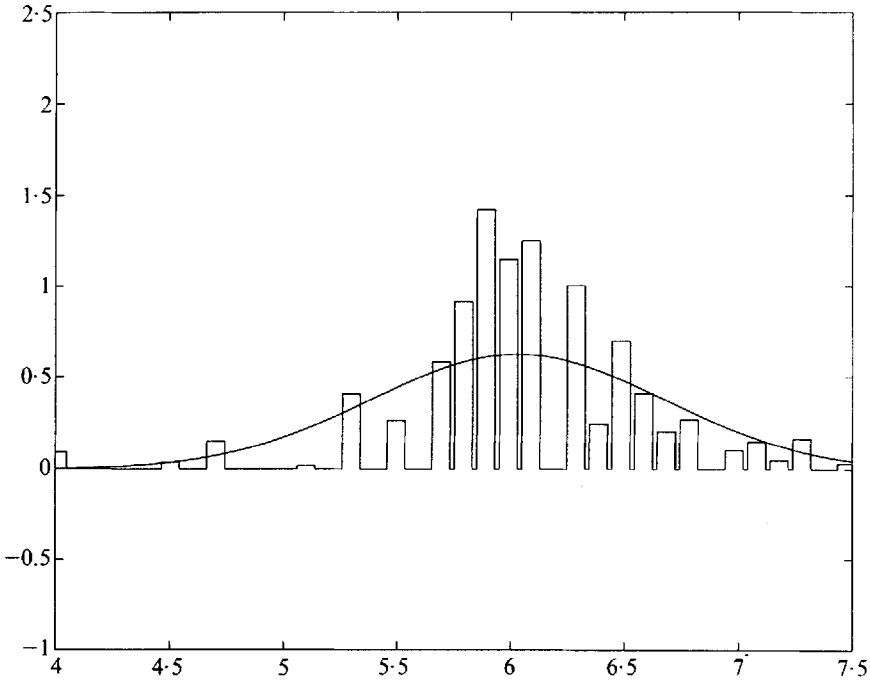


FIGURE 8
Multinomial and normal MLEs for $g_1^*$

they all lead to similar estimates of the local average treatment effect supports their credibility given the monotonicity and exclusion restrictions. The variance estimates, however, suggest that even in such a large sample it is difficult to obtain precise estimates of the higher order moments of the mixture components with weak instruments. This result agrees with the common wisdom that unless mixture models are appropriately restricted, their estimates can be unreliable. A further illustration of this point is that when we relax the restriction under the normal model the variances of $Y_i(0)$ and $Y_i(1)$ for compliers equal the variances of $Y_i(0)$ and $Y_i(1)$ for never-takers and always-takers, respectively, the estimates are outside the believable range: the variance of $Y(1)$ for compliers is estimated to be 0·02 and its mean 6·07—the distribution is concentrated around one of the minor modes of the sample distribution of $f_{11}(y)$.

## 7. A SMALL SIMULATION STUDY

In the previous section we presented new estimates for the local average treatment effect that differed considerably from the standard IV estimates with the AK data. To interpret these differences it is useful to see how these estimators perform in cases where we know the data generating distribution, and so we now present the results of a small simulation study. This is particularly important for the two estimators that can be viewed as based on parametrizations with many components, the nonnegative IV and the multinomial ML estimator where one might expect the small sample distributions to deviate considerably from the asymptotic distributions.

First we discuss the theoretical properties of the nonnegative IV and the multinomial ML estimators. Figure 9 displays the existence of a small sample bias towards ordinary

least squares regression estimates that may result from imposing the nonnegativity of the implicit density estimates. Suppose the outcomes $Y_i(0)$ and $Y_i(1)$ are binary. For ease of exposition assume that the probabilities of being a complier, never-taker or always-taker, $\phi_c$, $\phi_n$, and $\phi_a$ respectively, are positive and known. The figure plots the never-taker mean outcome $Y_i(0)$ vs. the directly estimable mean outcome for those assigned and receiving control, a group which is a mixture of compliers and never-takers with mixture proportion $\phi_c/(\phi_c+\phi_n)$. The region inside the parallelogram $\{(0,0), (\phi_c/(\phi_c+\phi_n), 0), (\phi_n/(\phi_c+\phi_n), 1), (1,1)\}$, corresponds to the set of $(E[Y_i|Z_i=0, D_i=0], E[Y_i|Z_i=1, D_i=0])$ consistent with a value of $E[Y_i(0)|C_i=c)]$ between zero and one.

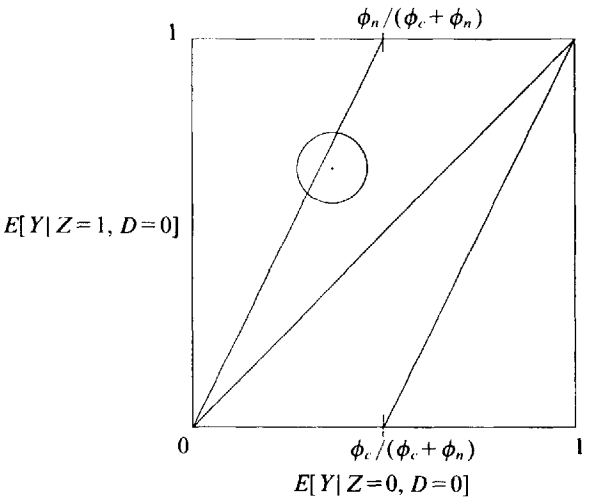


FIGURE 9

The dot and circle in Figure 9 denote the centre and contour of the sampling distribution of the unbiased moment estimates $\bar{Y}_{00}$ and $\bar{Y}_{10}$ of $E[Y(0)|Z=0, D=0]$ and $E[Y(0)|Z=1, D=0]$ respectively. As depicted, there is some probability mass of this distribution in the region where the implicit estimate of $E[Y_i(0)|C_i=c]=((\phi_c+\phi_n)/\phi_c)E[Y_{\text{obs},i}|Z_{\text{obs},i}=0, D_{\text{obs},i}=0]-(\phi_n/\phi_c)E[Y_{\text{obs},i}|Z_{\text{obs},i}=1, D_{\text{obs},i}=0]$ is negative: the area to the top/left of the line going through the origin and the point $(\phi_n/(\phi_c+\phi_n), 1)$. Both the nonnegative IV estimates and the multinomial and normal ML estimates of $E[Y_{\text{obs},i}|Z_{\text{obs},i}=0, D_{\text{obs},i}=0]$ and $E[Y_{\text{obs},i}|Z_{\text{obs},i}=1, D_{\text{obs},i}=0]$ by definition lie in the interior of the parallelogram $\{(0,0), (\phi_c/(\phi_c+\phi_n), 0), (\phi_n/(\phi_c+\phi_n), 1), (1,1)\}$, thereby biasing these estimates away from the unbiased moment estimates and towards the forty-five degree line where the mean of $Y_i(0)$ for compliers is the same as the mean of $Y_i(0)$ for never-takers. Combined with a similar bias in the estimates of $Y_i(1)$ for compliers towards equality of $E[Y_i(1)]$ for compliers and always-takers, this leads to a small sample bias of the estimates of the local average treatment effect towards the difference in outcomes by treatment, or the ordinary least squares estimates of the average treatment effect. At the same time, however, imposing these restrictions should lead to a reduction in the dispersion of the estimates. This is very similar to estimation in variance components models where unbiased estimators for the variances can lead to negative estimates: restricting the variance estimates in such models to be nonnegative leads to a reduction in mean squared error but also an increase in bias.

In the simulations each sample is of size 1000; 500 are randomly assigned $Z_i=0$ and 500 are randomly assigned $Z_i=1$. The population probability of being a complier is 0·1,

the probability of being a never-taker is 0·45 and the probability of being an always-taker is 0·45. The distributions of $Y_i(1)$ for always-takers and $Y_i(0)$ for never-takers are normal with mean zero and unit variance. The distribution of $Y_i(0)$ for compliers is normal with mean $-0·5$ and unit variance, and finally the distribution of $Y_i(1)$ for compliers is normal with mean 0·5 and unit variance. The local average treatment effect is $E[Y_i(1) - Y_i(0)|C_i = c] = 1·0$, and the population average treatment-control difference is $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = 0·1$. For the estimators based on histograms and multinomial distributions we use bins of width 0·6. For the normal ML estimates we impose, as with the actual data, equality of the variance of $Y(0)$ for compliers and never-takers and of the variance of $Y(1)$ for compliers and always-takers.

TABLE 3

*Estimates of the local average treatment effect (true value is* 1)

| | Bin width | Mean bias | rmse | Median bias | mae | 5th percentile | 95th percentile |
|---|---|---|---|---|---|---|---|
| OLS | | −0·90 | 0·90 | −0·90 | 0·90 | 0·00 | 0·21 |
| Standard IVE | | 0·17 | 0·92 | 0·05 | 0·47 | 0·05 | 2·05 |
| Normal MLE | | −0·01 | 0·62 | −0·00 | 0·39 | −0·04 | 2·05 |
| Nonnegative IVE | 0·6 | −0·31 | 0·53 | −0·31 | 0·37 | 0·01 | 1·42 |
| Multinomial MLE | 0·6 | −0·33 | 0·52 | −0·34 | 0·37 | 0·01 | 1·33 |

In Table 3 we present summary statistics (mean bias, root-mean-squared-error, median bias and median-absolute-error) over 500 replications for the four estimators described above and the OLS estimates, i.e. the treatment-control average difference. As expected, the nonnegative IV and multinomial ML estimator are biased towards the average treatment-control difference, but all three alternatives proposed in this paper have substantially lower rmse and somewhat lower median absolute error than the standard IV estimator. This partially reflects the thick tails of the standard IV estimator that are absent in the other estimators. The intuition for the thick tails of the standard instrumental variables estimator is clear: occasionally the moment estimate of denominator in the IV representation (9) is close to zero, suggesting the presence of few compliers. In that case the restrictions imply that the numerator has to be relatively small because few compliers can only lead to a relatively small average effect of $Z$ on $Y$. The standard IV estimator ignores this restriction and so can occasionally be very large.

Given the substantial bias of the multinomial maximum likelihood estimator, in practice it may be advisable to consider low dimensional parametrizations. Although the normal distribution used in the application and Monte Carlo investigation may be too limiting, generalizations to $t$-distributions or mixtures of normal distributions may be flexible enough to get close approximations to the four underlying outcome distributions while maintaining the advantages of low-dimensional parameterizations. These parametric models have the additional advantage that they are relatively easily extended to allow for covariates. If there is concern that the normal approximation to the maximum likelihood estimator is poor, Bayesian methods as described in Imbens and Rubin (1997) should be used.

## 8. CONCLUSION

In this paper we first show that with instrumental variables we can learn more from the data than just the average causal effect for those who are potentially affected by the instrument, the compliers: we can in fact estimate their entire outcome distributions under

both values of the treatment. These distributions may contribute to an understanding of the difference between simple treatment-control difference estimates of average causal effects and instrumental variables estimates, and can be helpful for policy purposes when there is concern about the distributional effects of programs.

Our second point is that conventional instrumental variables estimates are based on implicit estimates of density functions that are not restricted to be nonnegative. Because the assumptions underlying IV estimation, as explicated in AIR, restrict the distribution of the observable variables, they can be used to test the validity of the instrument even in the binary instrument, binary treatment case. Here we focused on the implications of the restrictions for estimation.

We also discussed three new methods for imposing nonnegativity on the density estimates. All three lead to similar inferences that are substantially different from that based on standard IV estimates in an example where we estimate the causal effect of education on earnings using quarter of birth as an instrument. This conclusion should be of concern to economists who routinely use these instrumental variables estimates, typically appealing to the lack of distributional and functional form assumptions as reasons to believe in their robustness. Two of the new methods are based on multinomial approximations to the four outcome distributions and the third relies on a normal approximation. The multinomial approximations show in simulations some bias towards the difference in average outcomes by treatment status estimates. The normal distribution based estimator performs very well in the simulations, outperforming the standard IV estimator, and giving credible answers with the actual AK data. Since this approach can easily be extended to allow for covariates and more general parametric models as well as for small sample Bayesian adjustments, we view it as the most attractive of the methods developed here.

*Acknowledgements.* The authors thank Joshua Angrist and Alan Krueger for making their data available, the National Science Foundation for financial support through grant SBR 9511718, the Alfred P. Sloan Foundation for a Sloan Research Fellowship for Imbens, and Joshua Angrist, Manuel Arellano, Gary Chamberlain and participants at the conference on programme evaluation at CEMFI (Madrid) for comments. An earlier version of this paper circulated under the title "On the Fragility of Instrumental Variables Estimator" as Harvard Institute of Economic Research Working paper no. 1675, February 1994.

REFERENCES

ANGRIST, J. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, **80**, 313–335.

ANGRIST, J. and KRUEGER, A. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, **106**, 979–1014.

ANGRIST, J. D. and IMBENS, G. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity", *Journal of the American Statistical Association*, **90**, 431–442.

ANGRIST, J., IMBENS, G. W. and RUBIN, D. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, **91**, 444–472.

BALKE, A. and PEARL, J. (1993), "Nonparametric Bounds of Causal Effects from partial Compliance Data", (Technical Report R-199, Computer Science Department, University of California, Los Angeles).

BJORKLUND, A. and MOFFITT, R. (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models", *Review of Economics and Statistics*, **54**, 42–49.

BUTCHER, K. and CASE, A. (1993). "The Effect of Sibling Composition on Women's Education and Earnings" (Unpublished Princeton University).

CARD, D. (1993), "Using Geographic Variations in College Proximity to Estimate the Returns to Schooling", (Working paper 4483, NBER).

CLEMENTS, N., HECKMAN, J. and SMITH, J. (1994), "Making the Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence from Randomized Trials with an Application to the National JTPA Experiment" (Mimeo, Department of Economics, University of Chicago).

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977), "Maximum Likelihood Estimation from Incomplete Data Using the EM Algorithm", *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

GRILICHES, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, **45**, 1–22.
GRONAU, R. (1974), "Wage comparisons: A Selectivity Bias", *Journal of Political Economy*, **82**, 1019–1043.
HECKMAN, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equations System", *Econometrica*, **46**, 931–961.
HECKMAN, J. and ROBB, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions", in Heckman, J., and Singer, B. (eds.), *Longitudinal Analysis of Labor Market Data* (New York: Cambridge University Press).
HECKMAN, J. (1990), "Varieties of Selectivity Bias", *American Economic Review*, Papers and Proceedings, 313–318.
HOLLAND, P. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association*, **81**, 945–970.
IMBENS, G. and ANGRIST, J. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, **62**, 467–476.
IMBENS, G. and RUBIN, D. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance", *Annals of Statistics*, **25**, 305 377.
KANE, T. and ROUSE, C. (1993), "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?", *Princeton University Industrial Relations Section*, Working Paper 311.
MADDALA, G. S. and NELSON, F. (1975), "Switching Regression Models with Exogenous and Endogenous Switching" *Proceeding of the American Statistical Association*, Business and Economics Section, 423–426.
McCLELLAN, M. and NEWHOUSE, J. (1994), "Does More Intensive Treatment of Acute Myocardial Infarctation in the Elderly Reduce Mortality", *Journal of the American Medical Association*, **272**, 859–866.
NEYMAN, J. (1923), "On the Application of Probability Theory to Agricultural Experiments Essay on Principles. Section 9", translated in *Statistical Science*, **5**, 465–480, 1990.
ROBINS, J. M. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies", in Sechrest, L. Freeman, H. and Bailey, A. (eds.), *Health Service Research Methodology: A Focus on AIDS* (vs. Public Health Service).
RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies", *Journal of Educational Psychology*, **66**, 688–701.
RUBIN, D. B. (1978), "Bayesian Inference for Causal Effects", *Annals of Statistics*, **6**, 34–58.
RUBIN, D. B. (1980), Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test", by Basu, *Journal of the American Statistical Association*, **75**, 591–593.
RUBIN, D. B. (1990), "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies", *Statistical Science*, **5**, 472–480.
WILLIS, R. and ROSEN, S. (1979), "Education and Self-Selection", *Journal of Political Economy*, **87**, 507–536.