Editor's Note. This article is the *JBES* 1999 Invited Lecture at the Joint Statistical Meetings, August 1999, in Baltimore, Maryland.

Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice

Joshua D. ANGRIST

Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142-1347, and National Bureau of Economic Research, Cambridge, MA (*angrist@mit.edu*)

Applied economists have long struggled with the question of how to accommodate binary endogenous regressors in models with binary and nonnegative outcomes. I argue here that much of the difficulty with limited dependent variables comes from a focus on structural parameters, such as index coefficients, instead of causal effects. Once the object of estimation is taken to be the causal effect of treatment, several simple strategies are available. These include conventional two-stage least squares, multiplicative models for conditional means, linear approximation of nonlinear causal models, models for distribution effects, and quantile regression with an endogenous binary regressor. The estimation strategies discussed in the article are illustrated by using multiple births to estimate the effect of childbearing on employment status and hours of work.

KEY WORDS: Instrumental variables; Labor supply; Sample selection; Semiparametric methods; Tobit.

Econometric models with dummy endogenous regressors capture the causal relationship between a binary regressor and an outcome variable. A canonical example is the evaluation of training programs, in which the binary regressor is an indicator for those who were trained and outcomes are earnings and employment status. Other examples include treatment effects in epidemiology and health economics, effects of union status, and the effects of teen childbearing on schooling or labor-market outcomes. All of these problems have a treatment-control flavor. The notion that treatment status is "endogenous" reflects the fact that simple comparisons of treated and untreated individuals are unlikely to have a causal interpretation. Instead, the dummy endogenous-variable model is meant to allow for the possibility of joint determination of outcomes and treatment status or omitted variables related to both treatment status and outcomes.

The principal challenge facing empirical researchers conducting studies of this type is identification. Successful identification in this context usually means finding an instrumental variable (IV) that affects outcomes solely through its impact on the binary regressor of interest. For better or worse, however, the formal discipline of econometrics is not much concerned with the "finding instruments problem"; this is a job left to the imagination of empirical researchers. This division of responsibility reminds me a little of Steve Martin's old joke about "how to make a million dollars and never pay taxes." First, Martin blandly suggests, "get a million dollars." In the same spirit, once you have solved the difficult problem of finding an instrument, then the tasks of estimation and inference—typically using two-stage least squares (2SLS)—look relatively straightforward.

But perhaps there is reason to worry about estimation and inference in this context after all. Even with a plausible instrument, the dummy endogenous-variables model still seems to raise some special econometric problems. For one thing, the endogenous regressor is binary, so perhaps a nonlinear first stage is in order. Second, and more importantly, in many cases the outcome of interest is also binary. Examples include employment status in the evaluation of training programs and survival status in health research. In other cases, the dependent variable has limited support, most often being nonnegative with a mass point at 0. Examples of this sort of outcome include earnings, hours worked, and expenditure on health care. The analysis of such limited dependent variables (LDV's) seems to call for nonlinear models like probit and tobit. This generates few stumbling blocks when the regressors are exogenous, but, with endogenous regressors, LDV models appear to present special challenges.

This article argues that difficulties with endogenous variables in nonlinear LDV models are usually more apparent than real. For binary endogenous regressors, at least, the technical challenges posed by LDV models come primarily from what I see as a counterproductive focus on structural parameters such as latent index coefficients or censored regression coefficients, instead of directly interpretable causal effects. In my view, the problem of causal inference with LDV's is not

fundamentally different from causal inference with continuous outcomes.

Section 1 begins by discussing identification strategies in LDV models with dummy endogenous regressors. I show that the auxiliary assumptions associated with structural modeling are largely unnecessary for causal inference. There is one important exception to this claim, however, and that is when attention focuses on conditional-on-positive effects, as in sample-selection models. For example, labor economists sometimes study the effect of an endogenous treatment on hours worked for those who work. Identification of such effects turns heavily on an underlying structural framework. On the other hand, the motivation for estimating this sort of effect is often unclear (to me). Moreover, claims for identification in this context typically strike me as overly ambitious because even an ideal randomized experiment fails to identify conditional-on-positive causal effects.

Setting aside the conceptual problems inherent in conditional-on-positive effects, a focus on causal effects instead of, say, censored regression parameters or latent index coefficients has a major practical payoff. First and most basic is the observation that if there are no covariates or the covariates are sparse and discrete, linear models and associated estimation techniques like 2SLS are no less appropriate for LDV's than for other kinds of dependent variables. This is because conditional expectation functions with discrete covariates can be parameterized as linear using a saturated model, regardless of the support of the dependent variable. Of course, relationships involving continuous covariates or a less-than-saturated parameterization for discrete covariates are usually nonlinear (even if the outcome variable has continuous support). In such cases, however, it still makes sense to ask whether nonlinear modeling strategies change inferences about causal effects.

If nonlinearity does seem important, it can be incorporated into models for conditional means using two new semiparametric estimators. The first, due to Mullahy (1997), is based on a multiplicative model that can be estimated using a simple nonlinear IV estimator. The second, developed by Abadie (1999), allows flexible nonlinear approximation of the causal response function of interest. In addition to new strategies for estimating effects on means, I also discuss estimates of the effect of treatment on distribution ordinates and quantiles using an approach developed by Abadie, Angrist, and Imbens (1998). This provides an alternative to the estimation of conditional-on-positive effects. Two advantages of these new approaches are their computational simplicity and weak identification requirements relative to other semiparametric approaches. Another advantage is the fact that they estimate causal effects directly and are not tied to a latent-index/censored-regression framework. The new estimators are illustrated by estimating the effect of childbearing on women's employment status and hours of work using multiple births as an instrument. This "twins instrument" was used to estimate the labor-supply consequences of childbearing by Rosenzweig and Wolpin (1980), Bronars and Grogger (1994), Gangadharan and Rosenbloom (1996), and Angrist and Evans (1998).

1. CAUSAL EFFECTS AND STRUCTURAL PARAMETERS

1.1 The Effects of Interest

The relationship between fertility and labor supply is of longstanding interest in labor economics and demography. For a recent discussion and references to the literature, see Angrist and Evans (1998), which is the basis of the empirical work in Section 4. The Angrist–Evans application is concerned with the effect of going from a family size of two children to more than two children. Let D_i be an indicator for women with more than two children in a sample of women with at least two children. The reasons for focusing on the transition from two to more than two are both practical and substantive. First, on the practical side, there are plausible instruments available for this fertility increment. Second, recent reductions in martial fertility have been concentrated in the 2–3 child range.

What is the object of interest in an application like this? Sometimes the purpose of research is merely descriptive, in which case we might simply compare the outcomes of women who have $D_i = 1$ with those of women who have $D_i = 0$. For this descriptive agenda, no special issues are raised by the fact that the dependent variable is limited, beyond the obvious consideration that if Y_i is binary, then one need only look at means. In contrast, if Y_i is a variable like earnings with a skewed distribution, the mean may not capture everything about labor-supply behavior that is of interest. In fact, a complete description would probably look at the entire distribution of outcomes, or at least at selected quantiles.

A major problem with descriptive analyses is that they may have little predictive value. Part of the motivation for studying labor supply and fertility is interest in how changes in government policy and the environment affect childbearing and labor supply. For example, we might be interested in the consequences of changes in contraceptive technology or costs, a motivation for studying the twins experiment mentioned by Rosenzweig and Wolpin (1980, p. 347). Similarly, one of the questions addressed in the labor-supply literature is to what extent declines in fertility have been a causal factor increasing female employment rates or reducing poverty. In contrast with descriptive analyses, causal relationships answer counterfactual questions and are therefore more likely to be of value for predicting the effects of changing policies or changing circumstances or understanding the past (e.g., see Manski 1996).

Causal relationships can be described most simply using explicit notation for counterfactuals or *potential outcomes*. This approach to causal inference goes back at least to R. A. Fisher, but the modern version is usually attributed to Rubin (1974, 1977). Let Y_{1i} denote the labor-market behavior of mother i if she has a third child, and let Y_{0i} denote labor-market behavior otherwise for the same mother. The average effect of childbearing on mothers who have a third child is

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]. \quad (1)$$

Note that the first term on the left side is observed, but the second term is an unobserved counterfactual average that we assume is meaningful.

The right side of (1) is often called the effect of treatment on the treated, and is widely discussed in the evaluation literature

(e.g., Rubin 1977; Heckman and Robb 1985; Angrist 1998). In the context of social-program evaluation, the effect of treatment on the treated tells us whether the program was beneficial for participants. This is not the only average effect of interest; we might also care about the unconditional average effect or the effect in some subpopulation defined by covariates (i.e., $E[Y_{1i} - Y_{0i}|X]$ for covariates, X). Ultimately, of course, we are also likely to want to extrapolate from the experiences of the treated to as-yet-untreated groups. Such extrapolation makes little sense, however, unless average causal effects in existing populations can be reliably assessed.

Simple comparisons of outcomes by D_i generally fail to identify causal effects. Rather, a comparison between treated and untreated individuals equals the effect of treatment on the treated plus a bias term:

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 1] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\quad = E[Y_{1i} - Y_{0i}|D_i = 1] \\ &\qquad + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\}. \end{aligned} \tag{2}$$

The bias term disappears in the childbearing example if childbearing is determined in a manner independent of a woman's potential labor-market behavior if she does not have children. In that case, $\{E[Y_{0i}|D_i = 0] - E[Y_{0i}|D_i = 1]\} = 0$, and simple comparisons identify the effect of treatment on the treated. But this independence assumption seems unrealistic because childbearing decisions are made in light of information about earnings potential and career plans.

1.2 Structural Models

What connects the causal parameters discussed in the previous section with the parameters in structural econometric models? Suppose that, instead of potential outcomes, we begin with a labor-supply model for hours worked, along the lines of many second-generation labor-supply studies (see Killingsworth 1983 for a survey). In this setting, childbearing is determined by comparing the utility of having a child and not having a child. We can model this process as

$$D_i = 1(X_i'\gamma > \eta_i), \tag{3a}$$

where X_i is a $K \times 1$ vector of observed characteristics that determine utility and η_i is an unobserved variable reflecting a person-specific utility contrast.

In a simple static model, labor supply is given by the combination of the participation decision and hours determination for workers. Workers chose their latent hours (y_i) by equating offered wages, w_i, with the marginal rate of substitution of goods for leisure, $m_i(y_i)$. Participation is determined by the relationship between w_i and the marginal rate of substitution at zero hours, $m_i(0)$. Since offered wages are unobserved for nonworkers and reservation wages are never observed, we decompose these variables into a linear function of observable characteristics and regression error terms (denoted v_{wi}, v_{mi}), as in the article by Heckman (1974) and many others:

$$w_i = X_i'\delta_w + \varphi_w D_i + v_{wi} \tag{3b}$$

and

$$m_i(y_i) = X_i'\delta_m + \psi y_i + \varphi_m D_i + v_{mi}. \tag{3c}$$

Equating (3b) and (3c) and relabeling parameters and the error term, we can solve for observed hours:

$$Y_i = X_i'\delta + \varphi D_i + \varepsilon_i \quad \text{if } w_i > m_i(0)$$

$$Y_i = 0 \text{ otherwise.}$$

Equivalently,

$$Y_i = 1(X_i'\delta + \varphi D_i > -\varepsilon_i)(X_i'\delta + \varphi D_i + \varepsilon_i). \tag{4}$$

Childbearing is said to be endogenous if the unobserved error determining D_i depends on the unobserved error in the participation and hours equations.

Since the structural equations tell us what a woman *would do* under alternative values of D_i, they describe the same sort of potential outcomes referred to in the previous model. The explicit link is

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i, \tag{5}$$

where

$$Y_{0i} = 1(X_i'\delta > -\varepsilon_i)(X_i'\delta + \varepsilon_i) \tag{6a}$$

and

$$Y_{1i} = 1(X_i'\delta + \varphi > -\varepsilon_i)(X_i'\delta + \varphi + \varepsilon_i). \tag{6b}$$

Once the structural parameters are known, we can use these relationships to write down expressions for causal effects. For example, the effect of treatment on the treated is

$$\begin{aligned} &E[Y_{1i} - Y_{0i}|D_i = 1] \\ &\quad = E\ 1(X_i'\delta + \varphi > -\varepsilon_i)(X_i'\delta + \varphi + \varepsilon_i) \\ &\qquad - 1(X_i'\delta > -\varepsilon_i)(X_i'\delta + \varepsilon_i)|X_i'\gamma > \eta_i\}. \end{aligned} \tag{7}$$

Note, however, that knowledge of the parameters on the right side of (7) is still not enough to evaluate this expression. The following lemma outlines the identification possibilities in this context:

Lemma. Assume that the covariates (X_i) are independent of continuously distributed latent errors, (η_i, ε_i). Then

1. If η_i is not independent of ε_i, and the probability of treatment is always nonzero, the effect of treatment on the treated is not identified without further assumptions (Heckman 1990).

2. If η_i is independent of ε_i, the effect of treatment on the treated is identified (exogenous treatment).

3. Suppose there is a covariate, denoted Z_i, with coefficient γ_1 in (3a), which is excluded from (3b) and (3c). Without loss of generality, assume $\gamma_1 > 0$. Then the local average treatment effect (LATE) given by $E[Y_{1i} - Y_{0i} | X_i'\gamma + \gamma_1 > \eta_i > X_i'\gamma]$ is identified (Imbens and Angrist 1994).

4. Suppose that LATE is identified as in 3 and that $P[D_i = 1 | Z_i = 0] = 0$. Then LATE equals the effect of treatment on the treated, $E[Y_{1i} - Y_{0i} | D_i = 1]$. Similarly, if $P[D_i = 1 | Z_i = 1] = 1$, LATE equals $E[Y_{1i} - Y_{0i} | D_i = 0]$ (Angrist and Imbens 1991).

This set of results can easily be summarized using nontechnical language. First, without additional assumptions, the effect of treatment on the treated is not identified in latent-index models. Second, the three positive identification results in the lemma, for exogeneous treatment, the LATE result, and the specialization of LATE to effects on the treated or nontreated, require no information about the structural model other than the distribution of D and Z. On the other hand, the LATE result for endogenous treatments in part 3 does not generally refer to the effect of treatment on the treated, so here the role played by the structural model in identifying causal effects merits further discussion.

The treatment effect captures the effect of treatment on the treated for those whose treatment status is changed by the instrument, Z_i. The data are informative about the effect of treatment on these people because the instrument changes their behavior. Thus, an exclusion restriction is enough to identify causal effects for a group directly affected by the "experiment" at hand. (Angrist, Imbens, and Rubin 1996 called these people *compliers*.) In some cases this is the set of all treated individuals, while in other cases this is only a subset. In any case, however, this result provides a foundation for credible causal inference because the assumptions needed for this narrow "identification-in-principle" can be separated from modeling assumptions required for smoothing and extrapolation to other groups of interest. The quality of this extrapolation is, of course, an open question and undoubtedly differs across applications. In my experience, however, often estimates of LATE differ little from estimates based on the stronger assumptions invoked to identify effects on the entire treated population. (Of course, I have to admit that, when simple and sophisticated estimation strategies do differ, I invariably prefer the simple! For an illustration of why this is, see the example that follows.)

Although extrapolation is an important part of what empirical researchers do, parameters like LATE and the effect of treatment on the treated provide a minimum-controversy jumping-off point for inference and prediction. Structural parameters are sometimes more closely linked to economic theory than average causal effects. But the ultimate goal of theory-motivated structural estimation seems to differ little from the causal agenda. For example, Keane and Wolpin (1997, p. 111) used structural models to "forecast the behavior of agents given any change in the state of the world that can be characterized as a change in their constraints." A prerequisite for this is credible assessment of causal effects of past changes. Structural parameters that are not linked to causal effects are not useful for this basic purpose [Mullahy (1998) made a similar point]. The rest of my discussion is therefore limited to models in which the effects of interest are defined directly in terms of potential outcomes.

2. CAUSAL EFFECTS ON LDV's

2.1 Average Effects in Experimental Data

Does the fact that a dependent variable is binary or nonnegative have any implications for empirical causal analysis? A useful starting point for this discussion is the analysis of randomized experiments, since some of the issues raised by the presence of LDV's have nothing to do with endogeneity. Suppose that D_i was randomly assigned, or at least assigned by some mechanism that ensures independence between D_i and Y_{0i}. In this case, a simple difference in means between those with $D_i = 1$ and with $D_i = 0$ identifies the effect of treatment on the treated:

$$\begin{aligned} &E[Y_i | D_i = 1] - E[Y_i | D_i = 0] \\ &\quad = E[Y_{1i} | D_i = 1] - E[Y_{0i} | D_i = 0] \\ &\quad = E[Y_{1i} | D_i = 1] - E[Y_{0i} | D_i = 1] \\ &\qquad \text{(by independence of } Y_{0i} \text{ and } D_i) \\ &\quad = E[Y_{1i} - Y_{0i} | D_i = 1] \\ &\qquad \text{(by linearity of conditional means).} \end{aligned} \tag{8}$$

If D_i is also independent of Y_{1i}, as would be likely in an experiment, then $E[Y_{1i} - Y_{0i} | D_i = 1] = E[Y_{1i} - Y_{0i}]$, the unconditional average treatment effect. (Usually this "unconditional" average still refers to a subpopulation eligible to participate in the experiment.)

Equation (8) shows that the estimation of causal effects in experiments presents no special challenges whether Y_i is binary, nonnegative, or continuously distributed. If Y_i is binary, then the difference in means on the left side of (8) corresponds to a difference in probabilities, while if Y_i has a mass point at 0, the difference in means is the difference in $E[Y_i | Y_i > 0, D_i] P[Y_i > 0 | D_i]$. But these facts have no bearing on the causal interpretation of estimates or, in the absence of further assumptions or restrictions, the choice of estimators.

2.2 Conditional-on-Positive Effects

In many studies with nonnegative dependent variables, researchers are interested in effects in a subset of the population with positive outcomes. Interest in conditional-on-positive effects is sometimes motivated by the following decomposition of differences in means, discussed by McDonald and Moffit (1980):

$$\begin{aligned} &E[Y_i | D_i = 1] - E[Y_i | D_i = 0] \\ &\quad = \{P[Y_i > 0 | D_i = 1] - P[Y > 0 | D_i = 0]\} \\ &\qquad \times E[Y_i | Y_i > 0, D_i = 1] \\ &\qquad + \{E[Y_i | Y_i > 0, D_i = 1] - E[Y_i | Y_i > 0, D_i = 0]\} \\ &\qquad \times P[Y > 0 | D_i = 0]. \end{aligned} \tag{9}$$

This decomposition describes how much of the overall treatment-control difference is due to participation effects (i.e., the impact on $1[Y_i > 0]$) and how much is due to an increase in intensity for those with $Y_i > 0$. For a recent example of this distinction, see Evans, Farrelly, and Montgomery (1999), who analyzed the impact of workplace smoking restrictions on smoking participation and intensity.

In an experimental setting, the interpretation of the first part of (9) as giving the causal effect of treatment on participation is straightforward. Does the conditional-on-positive difference in the second part also have a straightforward interpretation? The large literature constrasting two-part and sample-selection models for LDV's suggests not. (See, for example, Duan, Manning, Morris, and Newhouse 1984; Hay and Olsen 1984; Hay, Leu, and Roher 1987; Leung and Yu 1996; Maddala 1985; Manning, Duan, and Rogers 1987; Mullahy 1998.)

To analyze the conditional-on-positive comparison further, it is useful to write the mean difference by treatment status as follows (still assuming Y_{0i} and D_i are independent):

$$\begin{aligned} &E[Y_i|Y_i > 0, D_i = 1] - E[Y_i|Y_i > 0, D_i = 0] \\ &\quad = E[Y_{1i}|Y_{1i} > 0, D_i = 1] - E[Y_{0i}|Y_{0i} > 0, D_i = 0] \\ &\quad = E[Y_{1i}|Y_{1i} > 0, D_i = 1] - E[Y_{0i}|Y_{0i} > 0, D_i = 1] \\ &\quad = E[Y_{1i} - Y_{0i}|Y_{1i} > 0, D_i = 1] \qquad (10a) \\ &\qquad + \{E[Y_{0i}|Y_{1i} > 0, D_i = 1] - E[Y_{0i} > 0, D_i = 1]\}. \qquad (10b) \end{aligned}$$

On one hand, (10a) suggests that the conditional contrast estimates a potentially interesting effect, since this clearly amounts to a statement about the impact of treatment on the distribution of potential outcomes (in fact, this is something like a comparison of hazard rates). On the other hand, from (10b), it is clear that a conditional-on-working comparison does not tell us how much of the overall treatment effect is due to an increase in work *among treated workers*. The problem is that the conditional contrast involves different groups of people—those with $Y_{1i} > 0$ and those with $Y_{0i} > 0$. Suppose, for example, that the treatment effect is a positive constant, say, $Y_{1i} = Y_{0i} + \alpha$. Since the second term in (10b) must then be negative, the observed difference, $E[Y_i|Y_i > 0] - E[Y_i|Y_i > 0]$, is clearly less than the causal effect on treated workers, which is α in the constant-effects model. This is the selectivity-bias problem first noted by Gronau (1974) and Heckman (1974).

In principle, tobit and sample-selection models can be used to eliminate selectivity bias in conditional-on-positive comparisons. These models depict Y_i as the censored observation of an underlying continuously distributed latent variable. Suppose, for example, that

$$\begin{aligned} &Y_i = 1[Y_i > 0]Y_i^*, \text{ where} \\ &Y_i^* = Y_{01}^* + (Y_{1i}^* - Y_{0i}^*)D_i = Y_{0i}^* + D_i\alpha. \qquad (11) \end{aligned}$$

Recent studies with this type of censoring in a female labor-supply model are those of Blundell and Smith (1989) and Lee (1995), both of which include endogeneous regressors. Note that in this context the constant-effects causal model is applied to the latent variable, not the observed outcome.

Under a variety of distributional assumptions [e.g., normality, as in Heckman (1974) or weaker assumptions like symmetry, as in Powell (1986a)], the parameter α is identified. But what is the interpretation of α in a causal model? One possible answer is that α is the causal effect of D_i on Y_i^*, though Y_i^* is not observed, so this is not usually of intrinsic interest. However, a direct calculation using (11) shows that α is also a conditional-on-positive causal effect (for details, see the appendix). In particular,

$$\alpha = E[Y_{1i} - Y_{0i}|Y_{1i} > 0] \quad \text{if } \alpha < 0 \qquad (12a)$$

and

$$\alpha = E[Y_{1i} - Y_{0i}|Y_{1i} > 0] \quad \text{if } \alpha > 0. \qquad (12b)$$

Thus, the censored-regression model does succeed in separating causal effects from selection effects in conditional-on-positive comparisons. [We do not know a priori whether α is the effect in (12a) or (12b), but it seems reasonable to use the sign of the estimated α to decide.]

Although (11) provides an elegant resolution of the selection-bias dilemma, in practice I find the use of censored-regression models to accomplish this unattractive. One problem is conceptual. Although a censored-regression model seems natural for artificially censored data (e.g., topcoded variables in the Current Population Survey), the notion of a latent labor-supply equation that can take on negative values is less clear cut. The censoring in this case comes about because some people choose to work zero hours and not because of measurement problems [Maddala (1985) made a similar comment regarding Tobin's original application]. Here, an underlying structural model seems essential to the interpretation of empirical results. For example, in the labor-supply model from Section 1, the censored latent variable is the difference between unobserved offered wages and marginal rates of substitution. But even assuming the effect of regressors on this difference is of interest, the latent index coefficients alone have no predictive value for observable quantities.

Second, even if we adopt a theoretical framework that makes the latent structure meaningful, identification of a censored-regression model requires assumption beyond those needed for indentification of unconditional causal effects of the type described by (8). Semiparametric estimators that do not rely on distributional assumptions fail here because the regressor is discrete and there are no exclusion restrictions on the selection equation (see Chamberlain 1986). Moreover, in addition to requiring distributional assumptions, identification of α in (12) turns heavily on the additive, constant-effects model in (11). This is because $E[Y_{1i} - Y_{0i}|Y_{1i} > 0]$ involves the *joint* distribution of Y_1 and Y_0. The constant-effects assumption nails this joint distribution down, but the data contain information on marginal distributions only (which is why even randomized trials fail to answer the causal question that motivates samples-selection models).

These concerns, echoed by many applied researchers, stimulated investigation of alternative strategies for the analysis of nonnegative outcomes (e.g., see Moffit's 1999 survey).

The two-part model (2PM), introduced by Cragg (1971), and widely used in health economics, seems to provide a less demanding framework for the analysis of LDV's than sample-selection models. The two parts of the 2PM are $P[Y_i > 0|D_i]$ and $E[Y_i|Y_i > 0, D_i]$. Researchers using this model simply pick a functional form for each part. For example, probit or a linear probability model might be used for the first part and a linear or log-linear model might be used for the second part. Log-linearity for the second part may be desirable since this imposed nonnegativity of fitted values (e.g., see Mullahy 1998).

One attraction of the 2PM is that it fits a nonlinear functional form to the conditional expectation function (CEF) for LDV's even if both parts are linear. On the other hand, tobit-type sample-selection models fit a nonlinear CEF as well, provided there are covariates other than D_i. For example, the CEF implied by (11), with latent-index $Y_{0i}^* = X_i'\mu + D_i\alpha + \varepsilon_i$ and a Normal homoscedastic error, is

$$E[Y_i|X_i, D_i] = \Phi[(X_i'\mu + D_i\alpha)/\sigma][X_i'\mu + D_i\alpha] + \sigma\varphi[(X_i'\mu + D_i\alpha)/\sigma], \quad (13)$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the standard Normal density and distribution functions and σ is the standard deviation of ε_i (e.g., see McDonald and Moffit 1980). The derivative of this CEF with respect to D_i is $\Phi[(X_i'\mu + D_i\alpha)/\sigma]\alpha$.

The nonlinearity of (13) notwithstanding, at first blush the 2PM seems to provide a more flexible nonlinear specification than tobit or other sample-selection models. The latter imposes restrictions tied to the latent-index structure, while the two parts of the 2PM can be specified in whatever form seems convenient and fits the data well (a point made by Lin and Schmidt 1984). A signal feature of the 2PM, however, and the main point of contrast with sample-selection models, is that the 2PM does not attempt to solve the sample-selection problem in (10b). Thus, the second part of the 2PM does not have a clear-cut causal interpretation even if D_i is randomly assigned. Similarly, and of particular relevance here, is the point that instrumental variables that are valid for estimating the effect of D_i on Y_i are not valid for estimating the effect of D_i on Y_i conditional on $Y_i > 0$. The 2PM therefore seems ill suited for causal inference.

2.3 Effects on Distributions

Conditional-on-positive effects and sample-selection models are sometimes motivated by interest in the consequence of D_i beyond the impact on average outcomes [e.g., see Eichner, McClellan, and Wise's (1997) analysis of insurance effects on health expenditure]. Are there schemes for estimating effects on distributions that are less demanding than sample-selection models? Once the basic problem of identifying causal effects is resolved, the impact of D_i on the distribution of outcomes is identified and can be easily estimated.

To see this for the experimental (exogenous D_i) case, note that, given the assumed independence of D_i of Y_{0i}, the following relationship holds for any point, c, in the support of Y_i:

$$E[1(Y_i \le c)|D_i = 1] - E[1(Y_i \le c)|D_i = 0] = P[Y_{1i} \le c|D_i = 1] - P[Y_{0i} \le c|D_i = 1].$$

In fact, the entire marginal distributions of Y_{1i} and Y_{0i} are identified for those with $D_i = 1$. So it is easy to check whether D_i has an impact on the probability that $Y_i = 0$, as in the first part of the 2PM, or whether there is a change in the distribution of outcomes at any positive value. This information is enough to make social-welfare comparisons, as long as the comparisons of interest involve marginal distributions only.

2.4 Covariates and Nonlinearity

The conditional expectation of Y_i given D_i is inherently linear, as are other conditional relationships involving D_i alone. Suppose, however, that identification is based on a "selection-on-observables" assumption instead of presumed random assignment. This means that causal inference is based on the presumption that $Y_{0i} \perp\!\!\!\perp D_i|X_i$, and causal effects must be estimated after conditioning on X_i. For example, the effect of treatment on the treated can be expressed as

$$\begin{aligned} &E[Y_{1i} - Y_{0i}|D_i = 1] \\ &\quad = E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]|D_i = 1\} \\ &\quad = \int \{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]\} \\ &\qquad \times P(X_i = x|D = 1)dx, \end{aligned} \quad (14)$$

where $P(X = x|D = 1)$ is a distribution of density. Estimation using the sample analog of (14) is straightforward if X_i has discrete support with many observations per cell (e.g., see Angrist 1998). Otherwise, some sort of smoothing (modeling) is required to estimate the CEF's $E[Y_{1i}|X_i, D_i = 1]$ and $E[Y_{0i}|X_i, D_i = 0]$.

Regression provides a flexible and computationally attractive smoothing device. A conceptual justification for regression smoothing is that population regression coefficients provide the best (minimum mean squared error) linear approximation to $E[Y_i|X_i, D_i]$ (e.g., see Goldberger 1991). This "approximation property" holds regardless of the distribution of Y_i.

Separate regressions can be used to approximate $E[Y_{1i}|X_i, D_i = 1]$ and $E[Y_{0i}|X_i, D_i = 1]$, though this leaves the problem of estimating $P(X_i = x|D = 1)$ to compute the average difference in CEF's using (14). On the other hand, a simple additive model—say $E[Y_i|X_i, D_i] = X_i'\beta_r + \alpha_r D_i$—sometimes works well, in the sense that α_r—the "regression estimand"—is close to average effects derived from models that allow for nonlinearity and interactions between D_i and X_i. With discrete covariates and a saturated model for X_i, the additive model can be thought of as implicitly producing a weighted average of covariate-specific contrasts. Although the regression weighting scheme differs from that in (14), in practice, treatment-effect heterogeneity may be limited enough that alternative weighting schemes have little impact on the overall estimate (see Angrist and Krueger 1999 for more on this point).

3. ENDOGENOUS REGRESSORS: TRADITIONAL SOLUTIONS

LDV models with endogenous regressors were first estimated using distributional assumptions and maximum likelihood (ML). An early and influential work in this mold is that of Heckman (1978). This approach is not wedded to ML; Heckman (1978), Amemiya (1978, 1979), Newey (1986, 1987), and Blundell and Smith (1989) discussed two-step procedures, minimum-distance estimators, generalized least squares (GLS) estimators, and other variations on the traditional ML framework. Semiparametric estimators based on weaker distributional assumptions were discussed by, among others, Newey (1985) and Lee (1996).

The basic idea behind these strategies can be described as follows. Take the censored-regression model from (11) and add a latent first stage with instrumental variables, Z_i. So the complete model is

$$\begin{aligned} Y_i &= 1[Y_i > 0]Y_i^* \\ Y_i^* &= Y_{0i}^* + (Y_{1i} - Y_{0i}^*)D_i = Y_{0i}^* + D_i\alpha = \mu + D_i\alpha + \varepsilon_i \\ D_i &= 1[\gamma_0 + \gamma_1 Z_i > \eta_i]. \end{aligned} \tag{15}$$

The principal identifying assumption here is

$$(Y_{0i}, \eta_i) \perp\!\!\!\perp Z_i. \tag{16}$$

Parametric schemes use two-step estimators or ML to estimate α. Semiparametric estimators typically work by substituting an estimated conditional expectation, $\hat{E}[D_i|Z_i]$, for D_i and then using a nonparametric or semiparametric procedure to estimate the pseueo reduced-form [e.g., Manski's (1975) maximum score estimator for binary outcomes].

The problems I see with this approach are the same as listed for censored regression models with exogenous regressors. First, latent index coefficients are not causal effects. If the outcome is binary, semiparametric methods estimate scaled index coefficients and not average causal effects. Similarly, censored regression parameters alone are not enough to determine the causal effect of D_i on the observed Y_i. [I should note that this criticism does not apply to parametric estimators, where distributional assumptions can be used to recover causal effects, or to a recently developed semiparametric method by Blundell and Powell (1999) for continuous endogenous variables.] Second, this approach turns heavily on the latent-index setup. We can add to these points the fact that even weak distributional assumptions like conditional symmetry fail for the reduced-form error term, $(D_i - \hat{E}[D_i|Z_i]) - \varepsilon_i$, since D_i is binary (a point made by Lee 1996).

A final observation is that, given assumption (16), this whole setup is unnecessary for causal inference. The effect of treatment on the observed Y_i is identified for those women whose childbearing behavior is affected by the instrument. The twins instrument, for example, identifies the effect of D_i on mothers who would not have had a third child without a multiple second birth (see result 4 in the earlier lemma). It may be of interest to extrapolate from this group's experiences to those of other women, but the extrapolation problem is distinct from the problem of identifying the causal effect of childbearing in the "twins experiment."

4. NEW ECONOMETRIC METHODS

Conditional moments and other probability statements involving D_i alone are necessarily linear, but causal relationships involving covariates are likely to be nonlinear unless the covariates are discrete and the model is saturated. LDV models like probit and tobit are often used because of a concern that, unless the model is saturated, LDV's lead to nonlinear CEF's. The 2PM is sometimes also motivated this way (e.g., see Duan et al. 1984).

The headaches induced by nonlinearity notwithstanding, there are simple schemes for estimating causal effects in LDV models with endogenous regressors and covariates. In this section, I discuss three strategies for estimating effects on means and two for estimating effects on distributions. All but the first are based on new models and methods. None are tied to an underlying structural model.

The simplest option for estimating effects on means is undoubtedly to "punt" by using a linear, constant-effects model to describe the relationship of interest:

$$E[Y_{0i}|X_i] = X_i'\beta \tag{17a}$$

and

$$Y_{1i} = Y_{0i} + \alpha. \tag{17b}$$

The assumptions lead a linear causal model,

$$Y_i = X_i'\beta + \alpha D_i + \varepsilon_i, \tag{18}$$

easily estimated by 2SLS.

Although the constant-effects assumption is clearly unrealistic, in practice, more general estimation strategies often lead to similar average effects. A second issue that arises in this setting is that, because D_i is binary, a nonlinear first-stage such as probit or logit may seem appropriate for 2SLS estimation of (18). But the resulting second-stage estimates are inconsistent, unless the model for the first-stage CEF is actually correct. On the other hand, conventional 2SLS estimates using a linear probability model are consistent whether or not the first-stage CEF is linear. So it is generally safer to use a linear first-stage. Alternatively, consistent estimates can be obtained by using a linear or nonlinear estimate of $E[D_i|X_i, Z_i]$ *as an instrument.* (This is the same as the plug-in-fitted-values method when the first-stage is linear.) See Kelejian (1971) or Heckman (1978, pp. 946–947) for a discussion of this point and additional references. (It is also worth noting that a probit first-stage cannot even be estimated for the twins instrument because $P[D_i = 1|X_i, Z_i = 1] = 1$ for twins.)

4.1 IV for an Exponential Conditional Mean

A linear model like (18) is obviously unrealistic for binary outcomes and fails to incorporate natural restrictions on the CEF for nonnegative LDV's. This motivated Mullahy (1997) to estimate causal effects on nonnegative LDV's using a multiplicative model similar to that used by Wooldridge (1999) for panel data. The Mullahy (1997) model can be written in my notation as follows. Let X_i be a vector of observed covariates

as before, and let ω_i be an unobserved covariate correlated with D_i and Y_{0i}. The fact that this covariate is unobserved is the reason we need to instrument.

Let Z_i be a candidate instrument. Conditional on observed and unobserved covariates, both treatment status and the instrument are assumed to be independent of potential outcomes:

$$Y_{0i} \amalg (D_i, Z_i) \mid X_i, \omega_i \tag{19a}$$

Moreover, conditional on observed covariates, the candidate instrument is independent of the unobserved covariate:

$$Z_i \amalg \omega_i \mid X_i, \tag{19b}$$

though ω_i and D_i are not conditionally independent. The CEF for Y_{0i} is constrained to be nonnegative using an exponential model and (19a):

$$\begin{aligned} E[Y_{0i}|D_i, Z_i, X_i, \omega_i] &= \exp(X_i'\beta + \pi\omega_i) \\ &= \omega_i^* \exp(X_i'\beta), \end{aligned} \tag{19c}$$

where we also assume that the unobservable covariate has been defined so that $E[\omega_i^*|X_i] = 1$ (this is a normalization because we can define $\omega = \pi^{-1}[\upsilon - \ln(E[e^{\upsilon}|X])]$, where υ is unrestricted.)

Finally, the conditional-on-X-and-ω average treatment effect is assumed to be proportionally constant, again using an exponential model that ensures nonnegative fitted values:

$$\begin{aligned} E[Y_{1i}|D_i, Z_i, X_i, \omega_i] &= e^{\alpha} E[Y_{0i}|D_i, Z_i, X_i, \omega_i] \\ &= e^{\alpha} E[Y_{0i}|X_i, \omega_i]. \end{aligned} \tag{19d}$$

Combining (19c) and (19d), we can write $Y_i = \exp(X_i'\beta + \alpha D_i + \pi\omega_i) + \varepsilon_i$, where $E[\varepsilon_i|D_i, Z_i, X_i, \omega_i] \equiv 0$. These assumptions imply

$$E\{\exp(-X_i'\beta - \alpha D_i)Y_i - 1|X_i, Z_i\} = 0, \tag{20}$$

so (20) can be used for estimation provided Z_i has an impact on D_i. The proportional average treatment effect in this model is $e^{\alpha} - 1$, or approximately α for small values of α.

Estimation based on (20) guarantees nonnegative fitted values, without dropping zeros as a traditional log-linear regression model would. The price for this is a constant-proportional-effects setup and the need for nonlinear estimation. It is interesting to note, however, that with a binary instrument and no covariates (20) generates a simple closed-form solution for α. In the appendix, I show that without covariates the proportional treatment effect in Mullahy's model can be written

$$e^{\alpha} - 1 = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{-\{E[(1-D_i)Y_i|Z_i = 1] - E[(1-D_i)Y_i|Z_i = 0]\}}. \tag{21}$$

In light of this simplification, it seems worth asking if the right side of (21) has an interpretation that is not tied to the constant-proportional-effects model. To develop this interpretation, let D_{0i} and D_{1i} denote potential treatment assignments indexed against the binary instrument. For example, an assignment mechanism such as (15) determines D_{0i} as follows: $D_{0i} = 1[\gamma_0 > \eta_i]$ and $D_{1i} = 1[\gamma_0 + \gamma_1 > \eta_i]$. Using this notation, Imbens and Angrist (1994) showed that

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] \\ &\quad \times \{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]\}. \end{aligned} \tag{22}$$

The term $E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]$ is the LATE parameter mentioned in Lemma 1.

The same argument used to establish (22) can also be used to show a similar result for the average of Y_{0i} [i.e., instead of the average of $Y_{1i} - Y_{0i}$; see Abadie (2000a) for details]. In particular,

$$\begin{aligned} &E[(1 - D_i)Y_i|Z_i = 1] - E[(1 - D_i)Y_i|Z_i = 0] \\ &\quad = -E[Y_{0i}|D_{1i} > D_{0i}] \\ &\qquad \times \{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]\}. \end{aligned} \tag{23}$$

Substituting (22) and (23) for the numerator and denominator in (21), we have

$$e^{\alpha} - 1 = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]/E[Y_{0i}|D_{1i} > D_{0i}]. \tag{24}$$

Thus, Mullahy's procedure estimates a proportional LATE parameter in models with no covariates. The resulting estimates therefore have a causal interpretation under much weaker assumptions than (19a)–(19d). Moreover, the exponential model used for covariates in (19c) seems natural for nonnegative dependent variables and has a semiparametric flavor similar to proportional hazard models for duration data.

4.2 Approximating Causal Models

Suppose that the additive, constant-effects assumptions (17a) and (17b) do not hold but we estimate (18) by 2SLS anyway. It seems reasonable to imagine that the resulting 2SLS estimates can be interpreted as providing some sort of "best linear approximation" to an underlying nonlinear causal relationship, just as regression provides the best linear predictor (BLP) for any CEF. Perhaps surprisingly, however, 2SLS does not provide this sort of linear approximation in general. On the other hand, in a recent article Abadie (2000b) introduced a Causal-IV estimator that does have this property.

Causal-IV is based on the assumptions used by Imbens and Angrist (1994) to estimate average treatment effects. Under these assumptions, it can be shown that treatment is independent of potential outcomes conditional on being in the group whose treatment status is affected by the instrument (i.e, those with $D_{1i} > D_{0i}$, the group of "compliers" mentioned earlier). This independence can be expressed as

$$Y_{0i}, Y_{1i} \amalg D_i|X_i, D_{1i} > D_{0i}. \tag{25}$$

A consequence of (25) is that, for compliers, comparisons by treatment status have a causal interpretation:

$$\begin{aligned} &E[Y_i|X_i, D_i = 1, D_{1i} > D_{0i}] - E[Y_i|X_i, D_i = 0, D_{1i} > D_{0i}] \\ &\qquad = E[Y_{1i} - Y_{0i}|X_i, D_{1i} > D_{0i}]. \end{aligned}$$

For this reason, Abadie (2000b) called $E[Y_i|X_i, D_i, D_{1i} > D_{0i}]$ the Complier Causal Response Function (CCRF).

Now, consider choosing parameters b and a to minimize $E[(E[Y_i|X_i, D_i, D_{1i} > D_{0i}] - X_i'b - aD_i)^2|D_{1i} > D_0]$, or equivalently $E[(Y_i - X_i'b - aD_i)^2|D_{1i} > D_0]$. This choice of b and a provides the minimum mean squared error (MMSE) approximation to the CCRF. Since the set of compliers is not identified, this minimization problem is not feasible as written. However, it can be shown that

$$E[\kappa_i(E[Y_i|X_i, D_i, D_{1i} > D_{0i}] - X_i'b - aD_i)^2]/P[D_{1i} > D_{0i}] = E[(E[Y_i|X_i, D_i, D_{1i} > D_{0i}] - X_i'b - aD_i)^2|D_{1i} > D_0],$$

where $\kappa_i = 1 - D_i(1 - Z_i)/(1 - E[Z_i|X_i]) - (1 - D_i) \cdot Z_i/E[Z_i|X_i]$. Since κ_i can be estimated, the MMSE linear approximation to the CCRF can also be estimated. The result is a weighted least squares estimation problem with weights given by the estimated κ_i.

It is also worth noting that, although the preceding discussion focuses on linear approximation of the CCRF, any function can be used for the approximation. For binary outcomes, for example, we might use $\Phi[X_i'b + aD_i]$ and choose parameters to minimize $E[\kappa_i(Y_i - \Phi[X_i'b - aD_i])^2]$. Similarly, for nonnegative outcomes, it seems sensible to use an exponential model, $\exp[X_i'b + aD_i]$, and choose parameters to minimize $E[\kappa_i(Y_i - \exp[X_i'b - aD_i])^2]$. Abadie's framework allows flexible approximation of the CCRF using any functional form the researcher finds appealing and convenient. The resulting estimates have a robust causal interpretation, regardless of the shape of the actual CEF for potential outcomes.

4.3 Distribution and Quantile Treatment Effects

If Y_i has a mass point at 0, the conditional mean provides an incomplete picture of the causal impact of D_i on Y_i. We might like to know, for example, how much of the average effect is due to changes in participation and how much involves changes elsewhere in the distribution. This sometimes motivates separate analyses of participation and conditional-on-positive effects. In Section 2.3, I argued that questions regarding the effect of treatment on the distribution of outcomes are better addressed by comparing distributions. Simply comparing distributions is fine for the analysis of experimental data, but what if covariates are involved? As with the analysis of mean outcomes, the simplest strategy is 2SLS, in this case using linear probability models for distribution ordinates: $1[Y_i \le c] = X_i'\beta_c + \alpha_c D_i + \varepsilon_{ci}$. Of course, the linear model is not literally correct for conditional distribution except in special cases (e.g., a saturated regression parameterization).

Here too, the Abadie (2000b) weighting scheme can be used to generate estimates that provide an MMSE error approximation to the underlying distribution function (see Imbens and Rubin 1997 for a related approach to this problem). The estimator in this case chooses b_c and a_c to minimize the sample analog of the following population minimand:

$$E[\kappa_i(1[Y_i \le c] - X_i'b_c - a_c D_i)^2]. \tag{26}$$

The resulting estimates provide the BLP for $P[Y_i \le c|X_i, D_i, D_{1i} > D_{0i}]$. The latter quantity has a causal interpretation because

$$\begin{aligned} &P[Y_i \le c|X_i, D_i = 1, D_{1i} > D_{0i}] \\ &\quad - P[Y_i \le c|X_i, D_i = 0, D_{1i} > D_{0i}] \\ &= P[Y_{1i} \le c|X_i, D_{1i} > D_{0i}] \\ &\quad - P[Y_{0i} \le c|X_i, D_{1i} > D_{0i}]. \end{aligned}$$

Since the outcome is binary, nonlinear models such as probit or logit might also be used to approximate $P[Y_i \le c|X_i, D_i, D_{1i} > D_{0i}]$. Likewise, it is equally straightforward to use Abadie's weighting scheme to approximate the probability that the outcome falls into an interval instead of the cumulative distribution function.

An alternative to estimation based on (26) postulates a linear model for quantiles instead of distribution ordinates. Conventional quantile regression (QR) models for exogenous regressors begin with a linear specification: $Q_\theta[Y_i|X_i, D_i] = X_i'\mu_{\theta 0} + \mu_{\theta 1} D_i$. The parameters $(\mu_{\theta 0}, \mu_{\theta 1})$ can be shown to minimize $E[\rho_\theta(Y_i - X_i'm_0 - m_1 D_i.)]$, where $\rho_\theta(u_i) = \theta u_i^+ + (1-\theta)u_i^-$ is called the "check function" (see Koenker and Basset 1978). This minimization is computationally straightforward since it can be written as a linear programming problem.

The analysis of quantiles has two advantages. First, quantiles like the median, quartiles, and deciles provide benchmarks that can be used to summarize and compare conditional distributions for different outcomes. In contrast, the choice of c for the analysis of distribution ordinates is application-specific. Second, since nonnegative LDV's are often (virtually) continuously distributed away from any mass points, linear models are likely to be more accurate for conditional quantiles than for conditional probabilities. [For quantiles close to the censoring point, Powell's (1986b) censored QR model may be more appropriate.]

Abadie, Angrist, and Imbens (1998) developed a QR estimator for models with a binary endogeneous regressor. Their quantile treatment effects (QTE) procedure begins with a linear model for conditional quantiles *for compliers*: $Q_\theta[Y_i|X_i, D_i, D_{1i} > D_{0i}] = X_i'\beta_\theta + \alpha_\theta D_i$. The coefficient α_θ has a causal interpretation because D_i is independent of potential outcomes conditional on X_i, and the event $D_{1i} > D_{0i}$. Therefore, $\alpha_\theta = Q_\theta[Y_{1i}|X_i, D_{1i} > D_{0i}] - Q_\theta[Y_{0i}|X_i, D_{1i} > D_{0i}]$. In other words, α_θ is the difference in θ quantiles for compliers.

QTE parameters are estimated by minimizing a sample analog of the following weighted check-function minimand: $E[\kappa_i\rho_\theta(Y_i - X_i'b - aD_i)]$. As with the Causal-IV estimators, weighting by κ_i transforms the conventional QR minimand into a problem for compliers only. For computational reasons, however, it is useful to rewrite this as $E[\tilde{\kappa}_i\rho_\theta(Y_i - X_i'b - aD_i.)]$, where $\tilde{\kappa}_i = E[\kappa_i|X_i, D_i, Y_i]$. It is possible to show that $E[\kappa_i|X_i, D_i, Y_i] = P[D_{1i} > D_{0i}|X_i, D_i, Y_i] > 0$. This modified estimation problem has a linear programming representation similar to conventional quantile regression, since the weights are positive. Thus, QTE estimates can be computed using existing QR software, though this approach requires first-step

estimation of $\tilde{\kappa}_i$. In the following example, I use the fact that

$$\tilde{\kappa}_i = E[\kappa_i|X_i, D_i, Y_i] = 1 - \frac{D_i(1 - E[Z_i|Y_i, D_i, X_i])}{(1 - E[Z_i|X_i])} - \frac{(1 - D_i)E[Z_i|Y_i D_i, X_i]}{E[Z_i|X_i]} \quad (27)$$

and estimate $E[Z_i|Y_i, D_i, X_i]$ and $E[Z_i|X_i]$ with a probit first step. Since $\tilde{\kappa}_i$ is theoretically supposed to be positive, negative estimates of $\tilde{\kappa}_i$ are set to 0.

5. APPLICATION: LABOR–SUPPLY CONSEQUENCES OF A THIRD CHILD

The estimation uses a sample of roughly 250,000 married women aged 21–35 with at least two children drawn from the 1980 Census 5% file. About 53% of the women in this sample worked in 1979. Overall (i.e., including zeros), women in the sample worked about 17 hours per week. This can be seen in the first column of Table 1, which reports descriptive statistics and repeats some of the OLS and 2SLS estimates from Angrist and Evans (1998). Roughly 38% of women in this sample had a third child, an event indicated by the variable *Morekids*. The OLS estimates in column (2) show that women with *Morekids* = 1 were about 17 percentage points less likely to have worked in 1979 and worked about 6 hours fewer per week than women with *Morekids* = 0. The covariates in this regression are age, age at first birth, a dummy for male firstborn, a dummy for male secondborn, and Black, Hispanic, and other race indicators.

Table 1 also reports estimates of average effects computed using nonlinear models, still treating *Morekids* as exogenous. These average effects are approximations to effects of treatment on the treated, evaluated using derivatives to simplify computations. A detailed description of the average-effects calculations appears in the appendix. Probit estimates of the average impact on employment, shown in column (3), are almost identical to the OLS estimates. Similarly, the tobit estimate of the average effect of D_i on hours worked, shown in column (4), is −6.01, remarkably close to the OLS estimate of −6.02. (Note, however, that the *tobit coefficient* is −11.7.) Column (5) of Table 1 reports estimated average effects from a two-part model in which both parts of the model are linear. The 2PM estimate is virtually identical to the tobit and OLS estimates.

Roughly 8/10 of 1% of women in the extract had a twin second birth (multiple births are identified in the 1980 Census using age and quarter of birth). Reduced-form estimates of the effect of a twin birth are reported in columns (6) and (7). The reduced forms show that women who had a multiple birth were 63 percentage points more likely to have had a third child than women who had a singleton second birth. Mothers of twins were also 5.5 percentage points less likely to be working (standard error = .01) and worked 2.2 fewer hours per week (standard error = .37). The 2SLS estimates derived from these reduced forms, reported in column 8, show an impact of about −.09 (standard error = .02) on employment rates and −3.6 (standard error = .6) on weekly hours. These estimates are just over half as large as the OLS estimates, suggesting that the latter exaggerate the causal effects of childbearing. Of course, the twins instrument is not perfect, and the 2SLS estimates may also be biased. For example, twinning probabilities are slightly higher for certain demographic groups. But Angrist and Evans (1998) found that 2SLS estimates using twins instruments are largely insensitive to the inclusion of controls for mothers' personal characteristics.

Two variations on linear models generate estimates identical or almost identical to the conventional 2SLS estimates. This can be seen in columns (2) and (3) of Table 2, which report 2SLS estimates of a 2PM for hours worked and Abadie (2000b) Causal-IV estimates of linear models for employment and hours. The Causal-IV estimates use probit to estimate E[Z|X] and plug this into the formula for κ_i. The second step in Causal-IV estimation is a weighted least squares problem, with some negative weights. Since use of negative weights is nonstandard for statistical packages (e.g., Stata does not currently allow this), I used a MATLAB program available from Alberto Abadie to compute the estimates. The 2PM estimates in Table 3 were constructed from 2SLS estimates of a linear probability model for participation and 2SLS estimates of a linear model for hours worked conditional on working. In principle, the 2PM estimates do not have a causal interpretation because the instruments are not valid conditional on working. In practice, however, 2SLS estimates of the 2PM differ little from conventional 2SLS estimates.

The estimates of nonlinear/nonstructural models for mean effects are mostly similar to each other and to conventional

Table 1. Descriptive Statistics and Baseline Results

		Morekids exogenous				*Morekids endogenous*		
			Nonlinear models			*Reduced forms*		
Dependent variable	*Mean* (1)	*OLS* (2)	*Probit* (3)	*Tobit* (4)	*2PM* (5)	*(Morekids)* (6)	*(Dep.var)* (7)	*2SLS effect* (8)
Employment	.528	−.167	−.166	—	—	.627	−.055	−.088
	(.499)	(.002)	(.002)			(.003)	(.011)	(.017)
Hours worked	16.7	−6.02	—	−6.01	−5.97	.627	−2.23	−3.55
	(18.3)	(.074)		(.073)	(.073)	(.003)	(.371)	(.592)

NOTE: The sample includes 254,654 observations and is the same as that of Angrist and Evans (1998). The instrument is an indicator for multiple births. The mean of the endogenous regressor is .381. The probability of a multiple birth is .008. The model includes as covariates age, age at rst birth, boy rst, boy second, and race indicators. Standard deviations are shown in parentheses in column 1. Standard errors are shown in parentheses in other columns.

Table 2. Impact on Mean Outcomes (Morekids endogenous)

	Linear models			Nonlinear models			Structural models			
Dependent variable	2SLS (1)	2PM (2)	Causal-IV linear (3)	Mullahy (4)	Causal-IV probit (5)	Causal-IV expon. (6)	Bivar. probit (7)	Endog. tobit (8)	Mills ratio (9)	2SLS benchmark (10)
					A. With covariates					
Employment	−.088	—	−.089	—	−.088	—	−.124	—	—	−.089
	(.017)		(.017)		(.016)		(.016)			(.017)
Hours worked	−3.55	−3.54	−3.55	−3.82	—	−3.21	—	−3.81	−4.51	−3.60
	(.592)	(.598)	(.592)	(.598)		(.694)		(.580)	(.549)	(.599)
					B. No covariates					
Employment	−.084	—	−.084	—	−.084	—	−.086	—	—	−.084
	(.017)		(.017)		(.017)		(.017)			(.018)
Hours worked	−3.47	−3.37	−3.47	−3.10	—	−3.12	—	−3.35	−3.48	−3.52
	(.617)	(.614)	(.617)	(.561)		(.616)		(.642)	(.641)	(.624)

NOTE: Sample and covariates are the same as in Table 1. Results for nonlinear models are derivative-based approximations to effect on the treated. Causal-IV estimates are based on a procedure discussed by Abadie (2000b). Standard errors are shown in parentheses.

2SLS estimates. Results from nonlinear models are again reported as marginal effects that approximate average effects on the treated: For example, a causal probit model for employment status generates an average effect of −.088, identical (up to the reported accuracy) to the 2SLS estimate. Causal-IV estimation of an exponential model for hours worked, the result of a procedure that minimizes $E[\tilde{\kappa}_i(Y_i - \exp[X_i'b - aD_i])^2]$, generates an estimate of −3.21. This too differs little from the conventional 2SLS estimate of −3.55. Similarly, the Mullahy estimate of −3.82 in column (4) is less than 8% larger than conventional 2SLS in absolute value. It is noteworthy, however, that the Mullahy model generates results that change markedly (falling by about 20%) when the covariates are dropped. Since the covariates are not highly correlated with the twins instrument, this lack of robustness to the inclusion of covariates seems undesirable. On the other hand, without covariates, the Mullahy estimates are close to those from Causal-IV with an exponential model.

The bivariate probit estimate of the effect of childbearing on employment status, reported in column 7, is −.12, roughly a third larger in absolute value than the conventional 2SLS estimate. Interestingly, in another application, Abadie (2000b) also found that bivariate probit estimates are larger than Causal-IV. The gap between bivariate probit and Causal-IV estimates of effects on labor supply appears to be a consequence of the probit model for exogenous covariates. Without covariates, bivariate probit generates estimates that are very close to the results from the other estimators. It should be noted, however, that bivariate probit is not really appropriate for twins instruments because the probability $Morekids = 1$ is equal to 1 for twins. The probit ML estimator does not exist in this case. Therefore, to compute all of the estimates using a

Table 3. Impact on the Distribution of Hours Worked

	Distribution treatment effects						Quantile treatment effects		
	Exogenous Morekids			Endogenous Morekids					
Range (mean)	OLS (LPM) (1)	Probit (2)	Ordered probit (3)	2SLS (4)	Causal LPM (5)	Causal probit (6)	Quantile (value)	QR (7)	QTE (8)
0	.167	.166	.147	.088	.089	.088	.5	−8.92	−5.24
(.472)	(.002)	(.002)	(.002)	(.017)	(.017)	(.016)	(8)	(.186)	(.686)
1–10	.001	.001	.002	−.001	−.001	−.002	.6	−12.7	−7.98
(.046)	(.001)	(.001)	(.00003)	(.007)	(.007)	(.006)	(20)	(.172)	(.860)
11–20	−.015	−.014	−.011	.002	.002	.002	.7	−9.54	−6.19
(.093)	(.001)	(.001)	(.0001)	(.010)	(.010)	(.010)	(35)	(.184)	(1.07)
21–30	−.024	−.022	−.014	−.004	−.005	−.006	.75	−6.45	−3.60
(.075)	(.001)	(.001)	(.0002)	(.009)	(.009)	(.010)	(40)	(.156)	(1.17)
31–40	−.119	−.110	−.097	−.072	−.072	−.071	.8	−1.00	0.00
(.277)	(.002)	(.002)	(.001)	(.014)	(.014)	(.016)	(40)	(.286)	(1.09)
41+	−.009	−.008	−.023	−.009	−.009	−.006	.9	—	—
(.027)	(.001)	(.001)	(.0003)	(.005)	(.005)	(.007)	(40)		

NOTE: The table reports probability-model and (QTE) estimates of the impact of childbearing on the distribution of hours worked. The sample and covariates are the same as in Panel A of Table 2. Causal-IV estimates are based on a procedure discussed by Abadie (1999). Quantile treatments are based on a procedure discussed by Abadie et al. (1998). Standard errors are shown in parentheses. The standard errors in columns (7) and (8) are bootstrapped.

probit first-stage (bivariate probit, endogenous tobit, and Mills ratio), I randomly recoded 1% of D_i observations to 0. In principle, measurement error in a binary endogenous regressor biases 2SLS estimates (see Kane, Rouse, and Staiger 1999). In this case, however, column (10) shows that 2SLS estimates with the randomly recoded data differ little from 2SLS estimates using the original data.

Column (8) of Table 2 reports estimates of a structural tobit model with endogeneous regressors. These estimates were computed using a two-step estimator that approximates the MLE, again with recoded data. The two-step procedure adds a Mills-ratio type endogeneity correction to a censored regression and then applies tobit to the censored regression model with the correction term. [The correction term is $\rho\sigma_\varepsilon[D_i(-\varphi_i/\Phi_i)+(1-D_i)\varphi_i/(1-\Phi_i)]$, where ρ is the correlation between the latent error determining treatment assignment and the outcome residual; σ_ε is the standard deviation of the outcome residual and φ_i and Φ_i are Normal density and distribution functions evaluated at the probit first-stage fitted values; see Heckman and Robb (1985).]

As with the bivariate probit estimate, the endogenous tobit estimate is somewhat larger in magnitude than conventional 2SLS. This may be because of the Mills-ratio procedure for controlling for the endogeneity of D_i, more than the tobit correction for nonnegative outcomes. To see this, note that the Mills-ratio estimate ignoring censoring, reported in column (9), is considerably larger than the corresponding conventional 2SLS estimates. Interestingly, both the probit and tobit structural estimators generate results that are more sensitive to the inclusion of covariates than any of the other estimators except Mullahy's. In fact, Panel B of the table shows that, without covariates, all estimation techniques give very similar results. This is not surprising since, without covariates, the parametric assumptions used in these models are weaker.

The last set of results shows that childbearing is associated with marked changes in the distribution of hours worked, beyond the changes in participation already seen. This is apparent in the first columns of Table 3, which report the distribution of hours worked by interval, along with linear probability estimates of the relationship between childbearing and the probability of falling into each interval (these models include the same covariates used for Table 2). The largest entry in column (1) is for the probability of working zero hours. There is also a large negative effect on the probability of working 31–40 hours per week, which shows that women who have a third child are much less likely to work fulltime. Once again, probit average effects, reported in column 2, are almost indistinguishable from the corresponding OLS estimates. Estimates from an ordered probit model, reported in column (3), differ from OLS somewhat more but still generate a very similar pattern.

Like the 2SLS estimates for average outcomes, 2SLS estimates of linear probability models for the probability of falling into each interval show that models that treat childbearing as exogenous exaggerate the negative impact on labor supply. 2SLS estimates for the probability of working zero hours are identical (by construction) to those for employment in Table 1. The 2SLS estimates of the impact of childbearing on fulltime work are also considerably less than the corresponding OLS estimates.

Nonstructural Causal-IV models treating *Morekids* as endogenous generate estimates very close to 2SLS estimates for effects on the probabilities of hours falling into each interval. Columns (5) and (6) show that the results are also remarkably insensitive to whether a linear or probit model is used to approximate the distribution function. The estimates again indicate that childbearing changes the distribution of hours by raising the probability of nonparticipation and by reducing the probability of fulltime work.

Finally, the QTE estimator provides useful summary statistics for causal effects of childbearing on changes in the distribution of hours worked. These estimates were computed as the solution to a weighted quantile regression problem using (27) to construct weights, and the reported standard errors are from a bootstrap. Quantile regression estimates treating childbearing as exogenous show an estimated 9-hour decline in median hours worked, but the QTE estimator suggests that the causal effect of childbearing on median hours worked is only about five hours. Estimates at higher quantiles are similarly reduced when childbearing is treated as endogenous.

6. SUMMARY AND CONCLUSIONS

Structural parameters may be of theoretical interest but must ultimately be converted into causal effects if they are to be of use for policy evaluation or determining whether a trend association is causal. The problem of estimating causal effects for LDV's does not differ fundamentally from the analogous problem for continuously distributed outcomes. The key differences seem to me to be the increased likelihood of interest in distributional outcomes and the inherent nonlinearity of CEF's for LDV's in models with covariates. Without covariates, conventional 2SLS estimates capture both distributional effects and effects on means. Simple IV strategies developed by Mullahy (1997) and Abadie (2000b) can be used to estimate average effects in nonlinear models with covariates, while IV strategies for probability models and quantile regression can be used to estimate effects on distributions.

These approaches are illustrated here using twin births to estimate the labor-supply consequences of childbearing. Alternative nonstructural approaches to the estimation of causal effects using twins instruments generate similar average effects, whether or not the model is nonlinear. Structural estimates tend to be somewhat larger than nonstructural estimates when exogenous covariates are included, even though the covariates are not strongly related to the twins instrument. Since the structural models impose additional distributional and functional form assumptions, I see no reason to prefer them.

Finally, the various IV estimates of the effect of childbearing on the distribution of hours worked show that the impact of childbearing is characterized by substantially increased nonparticipation and by an almost equally large shift away from fulltime work. Estimates that treat childbearing as exogenous exaggerate the causal effect of childbearing on changes in distribution, as well as on average hours worked. These findings appear in results from both probability models and quantile models with endogenous regressors. Both types of models provide an interesting look at the impact of childbearing on

the distribution of hours worked, without the conceptual problems inherent in conditional-on-positive comparisons.

ACKNOWLEDGMENTS

I thank the editor and the session discussants for their comments. Special thanks go to Melissa Schettini for outstanding research assistance and Alberto Abadie for extensive comments and shared computer programs. Thanks also go to Daron Acemoglu, Bill Evans, Bill Greene, Guido Imbens, John Mullahy, Steve Pischke, and seminar participants at the University of Texas for helpful discussions and comments on an earlier draft. Data and computer programs used here are available on request.

APPENDIX

A.1 Derivation of Equation (12)

Drop the i subscripts. Note that

$$Y_0 = 1(Y_0^* > 0)Y_0^*$$
$$Y_1 = 1(Y_0^* + \alpha > 0)(Y_0^* + \alpha) = 1(Y_1^* > 0)Y_1^*.$$

Since D is independent of Y_0, and $Y_1 = 1(Y_0^* + \alpha > 0)(Y_0^* + \alpha)$, D is independent of Y_1. Conditional effects on the treated are therefore the same as conditional effects without conditioning on treatment status:

$$E[Y_1 - Y_0 | Y_1 > 0, D = 1] = E[Y_1 - Y_0 | Y_1 > 0]$$
$$E[Y_1 - Y_0 | Y_0 > 0, D = 1] = E[Y_1 - Y_0 | Y_0 > 0].$$

The averages on the right side are the causal effects of interest. The conditional expectations on the right side are evaluated as follows:

$$E[Y_1 | Y_1 > 0] = E[Y_1^* | Y_1^* > 0] = E[Y_0^* | Y_1^* > 0] + \alpha, \quad \text{(A.1)}$$

$$E[Y_0 | Y_1 > 0] = E[Y_0^* 1(Y_0^* > 0) | Y_1^* > 0]$$
$$= E[Y_0^* | Y_1^* > 0, Y_0^* > 0] P(Y_0^* > 0 | Y_1^* > 0).$$

If $(\alpha > 0) : Y_0^* > 0 \Rightarrow Y_1^* > 0$, this is $E[Y_0^* | Y_0^* > 0]$ $P(Y_0^* > 0 | Y_1^* > 0)$. If $(\alpha < 0) : Y_1^* > 0 \Rightarrow Y_0^* > 0$, so this is

$$E[Y_0^* | Y_1^* > 0]. \quad \text{(A.2)}$$

So $\alpha < 0 \Rightarrow E[Y_1 - Y_0 | Y_1 > 0] = \alpha$. Similarly,

$$E[Y_1 | Y_0 > 0] = E[1(Y_0^* + \alpha > 0)(Y_0^* + \alpha) | Y_0^* > 0]$$
$$= E[Y_0^* | Y_1^* > 0, Y_0^* > 0] P(Y_1^* > 0 | Y_0^* > 0)$$
$$+ \alpha P\ Y_1^* > 0 | Y_0^* > 0\ .$$

If $(\alpha > 0)$, $Y_0^* > 0 \Rightarrow Y_1^* > 0$, so

$$P(Y_1^* > 0 | Y_0^* > 0) = 1 \quad \text{(A.3)}$$

and $E[Y_0^* | Y_1^* > 0, Y_0^* > 0] = E[Y_0^* | Y_0^* > 0]$. Finally,

$$E[Y_0 | Y_0 > 0] = E[Y_0^* | Y_0^* > 0] \quad \text{(A.4)}$$

so $\alpha > 0 \Rightarrow E[Y_1 - Y_0 | Y_0 > 0] = \alpha$.

A.2 Derivation of Equation (21)

Drop the i subscripts. Note that $e^{-\alpha D} = [(1 - D) + De^{-\alpha}]$. Let $\beta^* = e^{-\beta}$ and $\alpha^* = e^{-\alpha}$. Z is binary and there are no covariates, so we can now write (20) as

$$\beta^* E[(1 - D)Y | Z = 1] + \beta^* \alpha^* E[DY | Z = 1] - 1 = 0 \quad \text{(A.5)}$$

and

$$\beta^* E[(1 - D)Y | Z = 0] + \beta^* \alpha^* E[DY | Z = 0] - 1 = 0. \quad \text{(A.6)}$$

Divided (A.5) by (A.6) to get rid of β^*, then solve for α^*. Subtract 1 and rearrange to get Equation (21).

A.3 Average Treatment Effects and Standard Errors for Nonlinear Models

Average treatment effects were calculated with the aid of derivative approximations so that all reported effects have the form "coefficient times scaling factor."

Probit. Note that $\Phi[X_i'\beta + \alpha] - \Phi[X_i'\beta]$ $\varphi[X_i'\beta + \alpha D_i] \cdot \alpha$, so the average effect on the treated can be approximated as $\{(1/N_1)\sum_i D_i \varphi[X_i'\beta + \alpha D_i]\} \cdot \alpha$, where $N_1 = \sum_i D_i$. This turns out to be accurate to three decimal places for the estimates in Table 1. Standard errors were calculated treating the scaling factor as nonrandom. This follows the convention for reporting marginal effects in programs like Stata; in practice, any correction for estimation of the scaling factor is likely to be minor. A similar approach was used for ordered probit.

Tobit. Tobit average treatment effects were approximated using a derivative formula that can be found, for example, in the work of Greene (1999): $E[Y_i | X_i, D_i = 1] - E[Y_i | X_i, D_i = 0]$ $\partial E[Y_i | X_i, D_i]/\partial D = \Phi[X_i'\beta + \alpha D_i] \cdot \alpha$. Average effects on the treated can therefore be approximated using $\{(1/N_1)\sum_i D_i \Phi[X_i'\beta + \alpha D_i]\} \cdot \alpha$. Standard errors were again calculated treating the scaling factor as nonrandom.

2PM. Let the part-1 coefficient be α_1 and the part-2 coefficient be α_2. Since the model multiplies parts and both parts are linear, the average effect is approximated using derivatives as $\alpha_1 E[Y_i | D_i = 1, Y_i > 0] + \alpha_2 P[D_i = 1 | Y > 0]$. Standard errors were calculated treating the scaling factors $E[Y | D_i = 1, Y_i > 0]$ and $P[D_i = 1 | Y > 0]$ as nonrandom, and using the fact that the estimates of α_1 and α_2 are uncorrelated.

Mullahy. Note that $E[Y_i | X_i, D_i, \omega_i^*] = \omega_i^* \exp[X_i'\beta + D_i\alpha]$, where this CEF has a causal interpretation. Again, using derivatives, we have $\omega_i^* \exp[X_i'\beta + \alpha] - \omega_i^* \exp[X_i'\beta]$ $\omega_i^* \exp[X_i'\beta + \alpha D_i] \cdot \alpha$. The model is such that $E[\omega_i^* | X_i]$ equals 1, but $E[\omega_i^* | X_i, D_i]$ is unrestricted. I ignore this problem and approximate the average effect on the treated as $\{(1/N_1)\sum_i D_i \exp[X_i'\beta + \alpha D_i]\} \cdot \alpha$. Standard errors were calculated treating the scaling factor as nonrandom.

Bivariate Probit. This is the same as probit, using parameters from the latent index equation for outcomes.

Endogenous Tobit. This is the same as tobit but using coefficients and predicted probability positive from the model with the compound Mills-ratio term included.

Mills Ratio. Standard errors were calculated treating the compound Mills-ratio term as known.

Causal-IV (nonlinear). Average effects were calculated as described previously for the probit and Mullahy (exponential) functional form. The first-stage estimates of $E[Z_i|X_i]$ needed to construct κ_i were estimated using probit. Standard errors for α were calculated using the same bootstrap procedure described for QTE. Abadie (2000b) also gave analytic formulas that take account of the first-step estimation of $E[Z_i|X_i]$. An assumption implicit in this scheme for reporting Causal-IV results is that it makes sense to convert conditional-on-X effects for compliers into overall average effects on the treated.

A.4 Computation of Quantile Treatment Effects and Standard Errors

QTE's were computed by plugging first-step estimates of $\tilde{\kappa}_i = E[\kappa_i|X_i, D_i, Y_i]$ into a weighted quantile regression calculation performed by Stata. Nonnegative estimates of $E[\kappa_i|X_i, D_i, Y_i]$ were constructed by separately estimating $E[Z_i|X_i]$ and $E[Z_i|Y_i, D_i, X_i]$ using probit and then trimming. In principle, standard errors should take account of this first-step estimation. An additional complication is that the analytic standard errors for QTE involve a conditional error density. In this case, I sidestepped messy analytic calculations by using a bootstrap procedure that repeats both the first-stage estimation of $\tilde{\kappa}_i$ and the second-step estimation of the parameters of interest in 100 replicate samples of 2,500 observations each. The 100 replicate samples were sampled without replacement using the Stata command *bsample*. The reported standard errors were calculated as $(N/N^*)^{1/2}bse_N$, where bse_N is the standard deviation of the 100 replicate estimates, $N = 2{,}500$, and N^* is the full sample size.

[Received October 1999. Revised July 2000.]

REFERENCES

Abadie, A. (2000a), "Bootstrap Tests for the Effect of a Treatment on the Distributions of an Outcome Variable," Technical Working Paper 261, National Bureau of Economic Research, Cambridge, MA.

——— (2000b) "Semiparametric Estimation of Instrumental Variable Models for Causal Effects," Technical Working Paper 260, National Bureau of Economic Research, Cambridge, MA.

Abadie, A., Angrist, J., and Imbens, G. (1998), "Instrumental Variables Estimation of Quantile Treatment Effects," Working Paper 229, National Bureau of Economic Research, Cambridge, MA.

Amemiya, T. (1978), "The Estimation of a Simultaneous Equation Generalized Probit Model," *Econometrica*, 46, 1193–1205.

——— (1979), "The Estimation of a Simultaneous Equation Tobit Model," *International Economic Review*, 20, 169–181.

Angrist, J. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applications," *Econometrica*, 66, 249–288.

Angrist, J., and Evans, W. (1998), "Children and Their Parents" Labor Supply: Evidence From Exogenous Variation in Family Size," *American Economic Review*, 88, 450–477.

Angrist, J., and Imbens, G. (1991), "Sources of Identifying Information in Evaluation Models," Technical Working Paper 117, National Bureau of Economic Research, Cambridge, MA.

Angrist, J., Imbens, G., and Rubin, B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J., and Krueger, A. (1999), "Empirical Strategies in Labor Economics," in *The Handbook of Labor Economics Volume IIIA*, eds. O. Ashenfelter and D. Card, Amsterdam: North-Holland, pp. 1277–1366.

Blundell, R. W., and Smith, R. J. (1989), "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models," *Review of Economic Studies*, 56, 37–58.

Blundell, R. W., and Powell, J. L. (1999), "Endogeneity in Single Index Models," mimeo, University College London, Dept. of Economics.

Bronars, S., and Grogger, J. (1994), "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment," *American Economic Review*, 84, 1141–1156.

Chamberlain, G. (1986), "Asymptotic Efficiency in Semi-Parametric Models With Censoring," *Journal of Econometrics*, 32, 189–218.

Cragg, J. G. (1971), "Some Statistical Models for Limited Dependent Variables With Application to the Demand for Durable Goods," *Econometrica*, 39, 829–844.

Duan, N., Manning, W. G., Jr., Morris, C. N., and Newhouse, J. P. (1984), "Choosing Between the Sample-Selection Model and the Multi-part Model," *Journal of Business & Economic Statistics*, 2, 283–289.

Eichner, M. J., McClellan, M. B., and Wise, D. A. (1997), "Health Expenditure Persistence and Feasibility of Medical Savings Accounts," in *Tax Policy and the Economy 11*, ed. J. Poterba, Cambridge, MA: National Bureau of Economic Research, pp. 91–128.

Evans, W., Farrelly, M. C., and Montgomery, E. (1999), "Do Workplace Smoking Bans Reduce Smoking?" *American Economic Review*, 89, 728–747.

Gangadharan, J., and Rosenbloom, J. L. (1996), "The Effects of Child-bearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment," Working Paper 5647, National Bureau of Economic Research, Cambridge, MA.

Goldberger, A. S. (1991), *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Greene, W. (1999), "Marginal Effects in the Censored Regression Model," *Economics Letters*, 64, 43–49.

Gronau, R. (1974), "Wage Comparisons—A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Hay, J. W., Leu, R., and Rohrer, P. (1987), "Ordinary Least Squares and Sample-Selection Models of Health-care Demand," *Journal of Business & Economic Statistics*, 5, 499–506.

Hay, J. W., and Olsen, R. J. (1984), "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," *Journal of Business & Economic Statistics*, 2, 279–282.

Heckman, J. J. (1974), "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

——— (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 931–960.

——— (1990), "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

Heckman, J. J., and Robb, R., Jr. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data* (Econometric Society Monographs Series, No. 10), eds. J. Heckman and B. Singer, Cambridge, U.K.: Cambridge University Press, pp. 156–246.

Imbens, G., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Imbens, G., and Rubin, D. B. (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

Kane, T. J., Rouse, C. E., and Staiger, D. (1999) "Estimating Returns to Schooling When Schooling Is Misreported," Working Paper W7235, National Bureau of Economic Research, Cambridge, MA.

Kelejian, H. H. (1971), "Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*, 66, 373–374.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Killingsworth, M. (1983), *Labor Supply*, Cambridge, U.K.: Cambridge University Press.

Keane, M. P., and Wolpin, K. (1997) "Introduction to the *JBES* Special Issue on Structural Estimation in Applied Microeconomics," *Journal of Business & Economic Statistics*, 15, 111–114.

Lee, M. J. (1995), "Semiparametric Estimation of Simultaneous Equations With Limited Dependent Variables: A Case Study of Female Labor Supply," *Journal of Applied Econometrics*, 10, 187–200.

——— (1996), "Nonparametric Two-Stage Estimation of Simultaneous Equations With Limited Endogenous Regressors," *Econometric Theory*, 12, 305–330.

Leung, S. F., and Yu, S. (1996), "On the Choice Between Sample Selection and Two-Part Models," *Journal of Econometrics*, 72, 197–229.

Lin, T.-F., and Schmidt, P. (1984), "A Test of the Tobit Specification Against an Alternative Suggested by Cragg," *Review of Economics and Statistics*, 66, 174–177.

Maddala, G. S. (1985), "A Survey of the Literature on Selectivity Bias As It Pertains to Health Care Markets," *Advances in Health Economics and Health Service Research*, 6, 3–18.

REFERENCES

Manning, W. G., Duan, N., and Rogers, W. H. (1987), "Monte Carlo Evidence on the Choice Between Sample Selection and Two-part Models," *Journal of Econometrics*, 35, 59–82.

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model," *Journal of Econometrics*, 3, 205–228.

——— (1996), "Learning About Treatment Effects From Experiments With Random Assignment of Treatments," *Journal of Human Resources*, 31, 709–733.

McDonald, J. F., and Moffitt, R. A. (1980), "The Uses of Tobit Analysis," *The Review of Economics and Statistics*, 62, 318–321.

Moffit, R. A. (1999), "New Developments in Econometric Methods for Labor Market Analysis," in *The Handbook of Labor Economics, Volume IIIA*, eds. O. Ashenfelter and D. Card, Amsterdam: North-Holland, pp. 1367–1398.

Mullahy, J. (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behaviour," *Review of Economics and Statistics*, 11, 586–593.

——— (1998), "Much Ado About Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics," *Journal of Health Economics*, 17, 247–281.

Newey, W. K. (1985), "Semiparametric Estimation of Limited Dependent Variable Models With Endogenous Explanatory Variables," *Annales de L'Insee*, 59/60, 219–237.

——— (1986) "Linear Instrumental Variable Estimation of Limited Dependent Variable Models With Endogenous Explanatory Variables," *Journal of Econometrics*, 32, 127–141.

——— (1987), "Efficient Estimation of Limited Dependent Variable Models With Endogenous Explanatory Variables," *Journal of Econometrics*, 36, 231–250.

Powell, J. L. (1986a), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, 1435–1460.

——— (1986b), "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

Rosenzweig, M. R., and Wolpin, K. (1980), "Life-Cycle Labor Supply and Fertility: Causal Inferences From Household Models," *Journal of Political Economy*, 88, 328–348.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1977), "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

Wooldridge, J. (1999), "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics*, 90, 77–97.

Comment: Binary Regressors in Nonlinear Panel-Data Models With Fixed Effects

Jinyong HAHN

Department of Economics, Brown University, Providence, RI 02912 (*jinyong_hahn@brown.edu*)

Angrist notes in his abstract that "much of the difficulty with limited dependent variables comes from a focus on structural parameters, such as index coefficients, instead of causal effects. Once the object of estimation is taken to be the causal effect of treatment, several simple strategies are available" (p. 2). I examine the consequence of such a perspective for inference in nonlinear panel-data models with fixed effects. I argue that (a) the "difficulty" indeed disappears sometimes and (b) structure of treatment assignment plays a crucial role for his strategy to be successful.

It is instructive to begin with a difficult nonlinear panel-data model with fixed effects. Consider a panel probit model with fixed effects:

$$\Pr[y_{it} = 1 | c_i, x_{i1}, x_{i2}] = \Phi(c_i + \theta\, x_{it}), \qquad i = 1, \ldots, n; \;\; t = 1, 2. \quad (1)$$

It is assumed that y_{i1} and y_{i2} are independent of each other given (c_i, x_{i1}, x_{i2}). Here, c_i denotes the unobserved fixed effects, and x_{it} denotes a binary treatment variable. We will assume that $(x_{i1}, x_{i2}) = (0, 1)$ for all i. The index coefficient of interest is θ. Therefore, traditional econometric analysis would focus on whether θ is identified and/ or $\sqrt{n}$-consistently estimable.

Estimation of the index coefficient θ is difficult for a number of reasons:

1. So far, no consistent estimator for θ has been developed for probit with a nonparametric specification of the conditional distribution of c_i given (x_{i1}, x_{i2}).

2. It is not even clear whether θ is identified or not. Manski's (1987) identification result requires infinite support for x_{it}, which cannot be satisfied due to the binary nature of x_{it}.

3. It is not clear whether the semiparametric information bound for θ is positive or not. It is quite possible that the information is actually 0. For example, Chamberlain (1992) showed that the information for θ is 0 for models with time dummies.

4. Chamberlain's (1984) conditional maximum likelihood estimator is applicable only to logit models.

Now, to assess whether changing the object of estimation simplifies statistical analysis, consider the average treatment effects, which in this particular case can be easily shown to be equal to

$$\beta = E[y_{i2} - y_{i1}]. \quad (2)$$

It is not difficult to see that a simple estimator $\hat{\beta} = 1/n \sum_{i=1}^{n} (y_{i2} - y_{i1})$ is $\sqrt{n}$-consistent. As Angrist argues, a focus on average causal effects dramatically reduces the difficulty of estimation. The secret is that $(x_{i1}, x_{i2}) = (0, 1)$ for all i, which effectively ensures that c_i is independent of (x_{i1}, x_{i2}). Difficulty of the index estimation listed previously is because

the independence between (x_{i1}, x_{i2}) and c_i cannot be exploited within the fixed-effects framework. Unfortunately, the additional information that (x_{i1}, x_{i2}) and c_i are independent of each other does not reduce difficulty in estimating the index θ: It is not yet clear whether θ can be $\sqrt{n}$-consistently estimated even with the random-effects assumption when the random effects are nonparametrically specified. On the other hand, estimating β becomes easier as a result of the constancy of the generalized propensity score $\Pr[x_{i1}, x_{i2}|c_i]$: Presence of unobserved c_i in the model was rendered irrelevant due to the constancy of the generalized propensity score. [See Imbens (1999) for a discussion on the generalized propensity score.]

It is interesting to note that, in the panel probit model (1), estimation of β is not necessarily simple unless the index structure is discarded altogether. It is useful to note that the new target parameter β could be estimated consistently using index structure *if* consistent estimators of θ and $\mathcal{L}$ are given. Here, $\mathcal{L}$ denotes the distribution of c_i. We may alternatively write (2) as

$$\int (\Phi(c+\theta) - \Phi(c))\, d\mathcal{L}(c), \tag{3}$$

which can in principle be estimated by using consistent estimators of θ and $\mathcal{L}$. Estimation of β using the alternative characterization (3) requires consistent estimation of an additional parameter $\mathcal{L}$, a parameter that was not given too much attention in the past. The problem is that not many consistent estimators of $\mathcal{L}$ are available. It is not yet clear whether the model satisfies the primitive conditions for consistency of the nonparametric maximum likelihood estimator (NPMLE) as discussed by Heckman and Singer (1984). The difficulty in estimating the target parameter using the expression (3), which is based on the index structure, is in sharp contrast to the ease of the estimation strategy using the expression (2), for which the index structure is irrelevant.

The preceding discussion suggests that the success of Angrist's perspective critically hinges on the structure of treatment assignment *and* careful reexpression of the new target parameter. If the joint distribution of c_i and (x_{i1}, x_{i2}) is completely unknown, it is clear that changing the target parameter does not ease the difficulty of estimation. Angrist's perspective therefore requires substantial effort in modeling such joint distribution. Whether such a modeling effort will be successful in dealing with nonlinear panel problems remains to be seen.

ADDITIONAL REFERENCES

Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics,* eds. Z. Griliches and M. D. Intriligator, Amsterdam: North-Holland, pp. 1247–1318.

——— (1992), "Binary Response Models for Panel Data: Identification and Information," unpublished manuscript.

Heckman, J., and Singer, B. (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271–320.

Imbens, G. (1999), "The Role of Propensity Score in Estimating Dose-Response Functions," Technical Working Paper 237, National Bureau of Economic Research, Cambridge, MA.

Manski, C. (1987), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica,* 55, 357–362.

Comment

Guido W. IMBENS

Department of Economics, University of California at Los Angeles, Los Angeles, CA 90095 (*imbens@econ.ucla.edu*)

It is a pleasure to comment on this article by Joshua Angrist, whose applications of instrumental-variables methods (Angrist 1989; Angrist and Krueger 1991) have been a source of inspiration for my own work in this area. As with Angrist's previous work on instrumental variables, the current article raises some controversial issues and makes a number of important points. Here I offer some comments on three of them. First, I shall discuss the issues raised in Section 1, "Causal Effects and Structural Parameters," concerning the goals of statistical inference. Angrist argues that many questions of interest are most easily formulated in terms of comparisons between realized and potential outcomes, the latter defined as outcomes that would have been observed under alternative states of nature. I shall explore some of the implications of this view for empirical practice and econometric theory. Second, I shall offer some remarks on the role of economic theory in specification and identification of econometric models, again reinforcing Angrist's point regarding the importance of formulating the key assumptions in terms of potential outcomes. Third, I shall discuss some of the issues related to the limited dependent nature of outcome variables for empirical practice, in particular in the presence of covariates. Partly motivated by the widespread perception of fundamental difficulties in applying instrumental-variables methods to data with limited dependent outcome variables, Angrist argues that standard linear model techniques are generally applicable. I agree with Angrist's position that most of these perceived problems are exaggerated but suggest that principled inference should nevertheless take account of the limited dependent nature of the outcome variables and use nonlinear models.

1. CAUSAL ESTIMANDS

In his textbook discussion of the difference between structural and reduced-form estimates, Goldberger (1997) wrote, following Marshak (1953), that the ultimate goal of

econometrics is to provide predictions. More specifically, in my view, the goal is to provide predictions of policy interventions. Using both economic theory and data, economists wish to inform policy discussions by providing predictions of states of the world under different policy choices. Based on comparisons of such predictions, policy makers can then choose among the different policies using some social welfare measure as objective function (e.g., Heckman and Smith 1997). Angrist argues that such questions are most easily formulated in terms of potential outcomes. Here I want to elaborate on that view.

Consider, as an example the problem faced by a policy maker contemplating a new tax in a market. To evaluate this policy, the policy maker wishes to take into account the effect of the tax on the quantity traded. Economic theory suggests that this effect depends on the slope of the supply and demand functions. The first step is therefore the estimation of these slopes, and in the remainder of this discussion I shall focus on this component of the policy-evaluation problem. In principle the policy maker may be interested in the entire distribution of the quantity traded under various taxes. Let us assume, however, that for purposes of evaluation of the policies it is sufficient to know the average effect of the policy on the quantity traded. If there are only two values for the policy—for example, no tax or a tax—the difference between these two averages is the key quantity of interest. Following Rubin (1974) I will refer to this as the *estimand.*

Note that the choice of estimand is distinct from the statistical question of the specification of the model. Often the statistical model is specified in such a way that a single parameter corresponds to the estimand. For example, in a structural interpretation of the linear regression model, the coefficients correspond to the effect of changing the covariates by a single unit. Such one-to-one correspondence, however, is the exception rather than the rule. Wooldridge (1992) made this point in the context of Box–Cox regression models. Such models are often used when a linear representation for $E[Y|X]$ is inappropriate. The Box–Cox regression model generalizes this linear form to $E[Y(\lambda)|X] = X'\beta$, where

$$Y(\lambda) = \begin{cases} (Y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \ln Y & \lambda = 0. \end{cases}$$

Although consistent estimators for β exist under these assumptions, Wooldridge stressed that because (a) the interpretation of β changes with the value of λ and (b) knowledge of β and λ is not sufficient for recovering $E[Y|X]$, there is no reason for economists to be interested in estimates of β under these assumptions. In other words, β cannot be the sole focus of the researcher because the question it answers changes with the value of nuisance parameters. Wooldridge then suggested an alternative specification that always allows the researcher to recover the conditional expectation $E[Y|X]$.

In empirical work this distinction between the estimand and the parameters of the statistical model is consistent with the now common practice of reporting estimates of average derivatives in binary response models rather than reporting estimates of the logit or probit coefficients. Unlike a linear regression model, there is no direct link from one of the coefficients in the logit or probit model to average causal effects, and thus there is no intrinsic interest in such coefficients.

This view is at odds, however, with a large part of the semiparametric literature. An exception is the work by Stoker (e.g., Stoker 1986), who focused on estimation of index coefficients in settings where these are proportional to average derivatives and thus directly linked to changes in predictions. Consider, for example, the work on semiparametric estimation of binary response models. In this literature, such models are estimated without making logistic or probit assumptions, instead only making conditional mean or median assumptions in a latent index interpretation (e.g., Manski 1985). This literature, however, has begged the question of why economists should be interested in the coefficient estimates in these models in the absence of a direct link between these coefficients and the choice probabilities or their derivatives. Similarly, some of the models with fixed effects in panel data with limited dependent variables have focused on estimation of parameters that in themselves do not allow for estimation of conditional expectations or their derivatives and thus do not allow for estimation of causal effects. See Arellano and Honoré (in press) for a survey of many of these methods.

2. IDENTIFICATION

After deciding on the estimand, the next step is to make substantive assumptions on the process that generated the data. This is where economic, as opposed to statistical, theory plays a key role. Theoretical considerations may suggest that certain variables have no direct causal effect on others because they do not enter into agents' utility function, nor do they affect the constraints these agents face. For example, in some markets it may be reasonable to postulate the existence of demand and supply function and assume that their intersection determines observed prices and quantities. In that case it may be argued that certain variables—for example, weather conditions in agricultural markets—affect supply at fixed prices but not demand because weather conditions do not affect utility of the buyers nor do they constrain their choices given prices. Similarly, theoretical considerations may suggest which variables, determine agents' fertility choices and which variables, are excluded from such choices, as in the structural models described in Section 1.2 of Angrist.

For the purpose of considering such exclusion restrictions, as well as other assumptions, it is important to formulate them in a way that economic theory can be brought to bear on them. This makes the formulation in terms of counterfactuals or potential outcomes that Angrist advocates particularly appropriate. The potential outcomes describe outcomes in different environments, and as such are the primitives of economic analyses, as well as choices under different sets of constraints, which are the result of agents solving constrained optimization problems. Since economic theory studies such optimization problems, it is therefore well equipped to assess assumptions formulated directly in terms of these potential outcomes. An example of the formulation of the critical assumptions in terms of such potential outcomes is Angrist, Imbens, and Rubin (1996, AIR from here on). In contrast, latent index models, although under some conditions mathematically equivalent to

the potential outcome framework (e.g., Vytlacil 1999), formulate the critical assumptions in terms of associations between observed variables and unobserved residuals, which appears more difficult to contemplate [see Imbens (1997) for a discussion of the confusion such formulations have caused in the statistics literature].

It is rare that economic theory is specific enough to determine the exact value of the estimand. More typical is that the theory is consistent with a range of values for the estimand. Observations on agents' choices and outcomes may be helpful in narrowing down this range. The econometrician's task is to link the data to the estimand. Typically a number of additional assumptions are made at this stage. Almost always it is assumed that there is only limited dependence, or no dependence at all, between choices made by different agents, and identification focuses on the link between the joint distribution of the observables, estimable in large samples, and the estimand. Two possibilities arise at this stage. Sometimes the estimand can be expressed as a functional of the joint distribution of the observables, in which case the estimand is identified. A leading example is where the estimand is the average treatment effect and theory suggests that assignment to treatment is random, or at least random conditional on a set of observed covariates (unconfounded assignment, selection on observables). Alternatively, the assumptions suggested by economic theory do not allow for the direct link between the distribution of observables and the estimand. In that case the researcher faces some choices. One option, advocated in a series of papers by Manski (see, for a general discussion, Manski 1995), is to estimate the range of values of the estimand consistent with the data given the substantive assumptions. Another option, followed in the current article by Angrist, is the local average-treatment-effect approach developed by Imbens and Angrist (1994) to consider what aspects of the estimand are identified given data and assumptions. In instrumental-variables settings, the population average treatment effect is often not identified, but the average effect for a specific subpopulation may be. In that case one may choose to estimate the average treatment effect for this subpopulation and leave the extrapolation to the principal estimand to the researcher, possibly aided by theoretical considerations. As Heckman wrote, "It is a great virtue of the LATE parameter that it makes the investigator stick to the data at hand, and separate out the aspects of an estimation that require out of sample extrapolation or theorizing from aspects of an estimation that are based on observable data" (Heckman 1999, p. 832).

Let us consider the case studied by Angrist, with its focus on the effect of having more than two children on labor supply. Angrist argues that the second birth being a multiple birth (e.g., twins) is a valid instrument for this effect. In terms of the AIR formulation, this requires a multiple birth to be as good as randomly assigned, and the absence of a systematic direct effect on labor supply other than through its effect on the number of children. Such assumptions may be controversial. For example, fertility treatments may lead to a systematic association between multiple births and choices made by couples, violating the first assumption. Even if we accept these assumptions, however, they only imply that the average causal effect of more kids on labor supply is identified for women who had a third child solely because their second birth was a multiple birth (compliers in the AIR terminology). In my view it is unlikely that this is the population of primary interest. Nevertheless, it is the only subpopulation the data are informative about in the sense of point identification under the substantive assumptions, and it would appear to offer some guidance regarding the population average causal effect to policy makers similar to the way in the medical world results from clinical trials in homogenous subpopulations are regarded as useful because they are viewed as indicative of population average causal effects.

3. LIMITED DEPENDENT VARIABLES

Typically economic theory offers some guidance concerning the determinants of certain outcomes without specifying the exact form or strength of their relationship. In that case statistical modeling is required to complete the specification. Consider the example Angrist studies with binary outcome, binary endogenous regressor, a binary instrument, and covariates. Angrist suggests as one possible approach estimating the average treatment effect through a linear probability model with instrumenting for an endogenous regressor. The benefits of the linear probability approach stemming from the linearity and robustness against misspecification of the first stage appear to me largely illusory. At this point the statistical modeling is only intended to provide flexible approximations to the underlying conditional distributions. This is a fundamentally different role from that played by the substantive assumptions that are essential for identification. Appeals to consistency under specific parameterizations therefore appear irrelevant—in a larger sample one may well wish to use a more flexible specification because less smoothing is required. In addition to finding the alleged benefits of the linear probability model unpersuasive, I find its disadvantages troubling. Within small subpopulations characterized by extreme values of the covariates, the smoothing implicit in linear probability models is likely to lead to unattractive predictions compared to predictions based on nonlinear models that respect the limited-dependent-variable nature of the outcomes.

An alternative approach is followed in the study of the effect of flu shots on hospitalization rates using randomized incentives for vaccination by Hirano, Imbens, Rubin, and Zhou (2000, HIRZ from here on). Given their assumptions, extensions of those made by AIR to the case with exogenous covariates, there are three subpopulations—compliers (units who change treatment status in response to a change in the value of the instrument), always-takers (who always take the treatment, irrespective of the value of the instrument), and never-takers (who never take the treatment, irrespective of the value of the instrument). HIRZ modeled the conditional distribution of these three "types" conditional on covariates as a trinomial distribution:

$$\Pr(\text{Type}_i = c | X_i = x) = \frac{\exp(x'\psi_c)}{1+\exp(x'\psi_c)+\exp(x'\psi_a)},$$

$$\Pr(\text{Type}_i = a | X_i = x) = \frac{\exp(x'\psi_a)}{1+\exp(x'\psi_c)+\exp(x'\psi_a)},$$

and

$$\Pr(\text{Type}_i = n | X_i = x) = 1 - \Pr(\text{Type}_i = c | X_i = x) - \Pr(\text{Type}_i = a | X_i = x).$$

Now compare this setup to the selection models Angrist describes in Section 3. In the selection models, the equation describing the endogenous regressor is $D_i = 1\{\gamma_0 + \gamma_1 Z_i + \gamma_2' X_i > \eta_i\}$. Suppose that the instrument is binary and that γ_1 is positive. Then the two models are very similar, with units with $\gamma_0 + \gamma_1 + \gamma_2' X_i > \eta_i$ in the selection model classified as always-takers in the potential outcome framework (because, irrespective of the value of the instrument, $D_i = 1$ for such units), units with $\gamma_0 + \gamma_2' X_i < \eta_i$ classified as never-takers (because, irrespective of the value of the instrument, $D_i = 0$ for such units), and the units with $\gamma_0 + \gamma_2' X_i < \eta_i < \gamma_0 + \gamma_1 + \gamma_2' X_i$ classified as compliers.

One advantage of the trinomial model is that it easily generalizes to provide an arbitrarily good fit to any conditional trinomial distribution by including higher-order terms and interactions in the covariates. If there are no substantive reasons to impose additional restrictions one should not impose them implicitly in the specification of the statistical model. In particular, in the selection model it is not sufficient to add higher-order terms to the covariate vector to provide an arbitrarily good fit to the trinomial distribution. Such an approximation would have to involve heteroscedasticity and other distributional extensions that are not straightforward to implement in the selection model.

Conditional on the individual's type, HIRZ specified the outcome distributions given covariates as logistic regression models. Again the aim is to provide a flexible approximation to the conditional distribution in a manner that does not impose any implicit restrictions. Given that for a binomial distribution the logistic regression model can be thought of as providing a linear approximation to the log odds ratio, this choice is again an appealing one. An alternative is the probit model, which also provides a good approximation. Less attractive here is the linear probability model since it requires inequality restrictions on the parameters if the implicit estimates of the probabilities are to be bounded between 0 and 1.

In cases with other limited dependent variables, alternative nonlinear models may be appropriate. For example, if the outcomes are durations, subject to censoring, models specified in terms of hazard functions (e.g., Lancaster 1979) may be convenient for dealing with such data.

ADDITIONAL REFERENCES

Angrist, J. (1989), "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 Ecomonics*, 106, 979–1014.

Arellano, M., and Honoré, B. (in press), "Panel Data," in *Handbook of Econometrics*, eds. J. Heckman and E. Leamer, Amsterdam: Elsevier, North-Holland.

Goldberger, A. (1991), *A Course in Econometrics*, Cambridge, MA: Cambridge University Press.

Heckman, J. (1999), "Instrumental Variables: Response to Angrist and Imbens." *Journal of Human Resources*, 34, 828–837.

Heckman, J., and Smith, J. (1997), "Evaluating the Welfare State," in *Econometrics and Economics in the 20th Century: The Ragnar Frisch Centenary*, ed. S. Strom, New York, Cambridge University Press, pp. 214–318.

Hirano, K., Imbens, G., Rubin, D., and Zhou, A. (2000), "Estimating the Effect of Flu Shots in a Randomized Encouragement Design," *Biostatistics*, 1, 69–88.

Imbens, G. (1997), Book Review of *The Foundations of Econometric Analysis*, by D. Hendry and M. Morgan, *Journal of Applied Econometrtics*, 12, 91–94.

Lancaster, T. (1979), "Econometric Methods for the Analysis of Duration Data," *Econometrica,*

Manski, C. (1985), "Semiparametric Analysis of Discrete Response," *Journal of Econometrics*, 27, 313–333.

——— (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.

Marshak, J. (1953), "Economic Measurements for Policy and Prediction," in *Studies in Econometric Method*, eds. W. Hood and T. Koopmans, New York, Wiley, pp. 1–26.

Stoker, T. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

Vytlacil, E. (1999), "Independence, Monotonicity, and Latent Index Models: An Equivalency Result," unpublished manuscript, University of Chicago, Dept. of Economics.

Wooldridge, J. (1992), "Some Alternatives to the Box–Cox Regression Model," *International Economic Review*, 33, 935–955.

Comment

Robert A. Moffitt

Department of Economics, Johns Hopkins University, Baltimore, MD 21218, and National Bureau of Economic Research, Cambridge, MA (*moffit@jhu.edu*)

The Problem. Although the article by Angrist ranges across a number of issues, much of the discussion, and the article title, suggests that the problem of concern is that instrumental variables (IV) cannot be used in one of three common models. Let the first model be $y = \alpha + \beta d + x\delta + \epsilon$, where y is an absolutely continuous variable but d is binary, and where x is independent of ϵ but d is not. Then β can be consistently estimated with IV (Heckman and Robb 1985). Let the second model be $y^* = \alpha + \beta d^* + x\delta + \epsilon$, where y^* and d^* are continuous and where $y = 1(y^* > 0)$ and $d = d^*$ are the observed variables. The parameters of this model can likewise be estimated by IV with some auxiliary assumptions (Newey 1986; see Blundell and Smith 1993 for a review of alternative methods). But let the third model be $y^* = \alpha + \beta d + x\delta + \epsilon$, where

again $y = 1(y^* > 0)$ but now d is binary. The parameters of this model cannot be estimated by standard IV. Many of the examples in the Angrist article involve the censored regression model rather than the binary choice model, but this and the other main points in the article apply to both.

The Angrist Solution. The Angrist solution to the problem posed by the third of these models, where IV is not applicable, is to declare β an uninteresting parameter and not worthy of estimation [or, in his words, to "punt" (p. 8)]. This is reminiscent of the solution to the Vietnam War suggested by some 1960s commentators: The United States should have simply declared victory, quickly withdrawn, and hoped that no one noticed that they had in fact lost.

If β is uninteresting, what is interesting, according to Angrist? His answer is that interest should center on least squares approximations with y as the dependent variable and d and x as regressors. If endogeneity of d is a problem, linear IV should be applied, with linear IV given a LATE interpretation. He provides an empirical illustration.

I will comment on three questions raised by the Angrist article: (1) Is β an interesting parameter? (2) Is the Linear IV estimator using observed y and d, which Imbens and Angrist (1994) relabeled the LATE, an interesting quantity? (3) Is the illustrative application given in the Angrist article interesting?

Is β an Interesting Parameter? As Angrist notes, many of the issues raised by his solution have nothing to do with the endogeneity of d. Whether β is an interesting parameter is one of these, at least in part. Given Angrist's preference for linear projections, his position necessarily implies that all nonlinear models, of which the latent index model (LIM) is just one example, are uninteresting (Angrist is not completely consistent in his position because he does propose a nonlinear model with an exponential conditional mean function in one section of his article, that model is subject to the same objections that he subjects the LIM to—arbitrary nonlinearities, etc.). Thus his position boils down to a preference, in the binary dependent-variable case, for the linear probability model (LPM) over the LIM and other nonlinear models. This issue has been discussed for many years, even in a simultaneous-equations context [e.g., Heckman and MaCurdy (1985), which Angrist does not reference].

The difficulty with Angrist's polar position on this issue—that the LPM is the only model of interest—is that it is fundamentally untenable. The LPM is attractive because it is easy to interpret, providing parameter estimates that do not require transformation to learn the effects of a regressor on the mean of the dependent variable. It has a role to play in empirical work in summarizing the data as regards the conditional mean function and for initial explorations of the data, and virtually all practitioners use it for this purpose. But beyond this it has no defense. Either it is equivalent to the latent-variable model if the model is saturated or, if the model is not saturated, imposes functional form restrictions that may not hold and may fit the data worse than the latent-variable model.

Thus, if the model is $y = 1(\alpha + \beta d + \epsilon > 0)$, where d is a single dummy variable regressor independent of ϵ, the probit (say) estimates of α and β map one-to-one onto the least squares intercept and coefficient on d in the LPM, and hence there is no gain to estimating the model either way because the estimators are equivalent. The same point extends to the case of a multinomial d. If the model is $y = 1(\alpha + \beta d + x\delta + \epsilon > 0)$, where x and δ are vectors of exogenous covariates and coefficients, respectively, the LPM estimate of the effect of d on $E(y)$ conditional on x may be demonstrably worse than the probit estimator of the same effect, even for the object of estimating $E(y|d, x)$ at points $\{d, x\}$ observed in the data (Angrist falls back on the weak principle that least squares estimates provide best linear approximations, but the issue is that the true model might be nonlinear). Nothing in Angrist's article provides evidence that the probitor any other latent index formulation provides an inferior estimate of $[E(y|d_1, x) - E(y|d_0, x)]$ compared to the LPM, for the two estimators smooth the joint response function with respect to d and x in different ways.

Although the latent-variable model has no necessary claim for superiority in the class of all nonlinear models, its popularity rests on its ability to generate a wide variety of nonlinearities with a relatively parsimonious specification, arising from the convolution of the latent index with the cdf of ϵ. Expansion of the LPM to incorporate equivalent nonlinearities is cumbersome and inefficient.

The possible inferiority of the LPM in capturing nonlinearities in nonsaturated models is also important for interpolation and extrapolation to points not in the observed data. Getting the nonlinearities right in the observed data is important in interpolation and extrapolation if the true model is nonlinear, as the binary choice model necessarily is (since y is bounded by 0 and 1). Although Angrist at one point does mention prediction—stating that the linearized models he proposes are a good "jumping-off point for any prediction exercise," a statement without support—he is, for the most part, not interested in prediction so much as summarizing the observed data. This is fine but, unfortunately, prediction outside the observed, historical data is the pervasive concern of the policy makers who are the ultimate consumers of applied economic research, for they must make predictions of new policies in new environments on a daily basis as part of their jobs. The methods proposed in Angrist's article are therefore not very useful for the formulation of new policy.

It should be noted that some make the argument that theory and economic models must play a role in guiding the formulation of empirical relationships used for prediction outside the range of the observed data. That position is not taken here because it is not necessary to establish the potential superiority of the LIM over the LPM in nonsaturated models and for prediction; Occam's razor makes it unnecessary.

It should also be noted that these issues have nothing to do with causal effects because d is assumed exogenous. Thus the opposition that Angrist poses between LIM's and "causal effects" models is a false one, at least in general. If a "causal effect" is defined as the true effect of a variable d on only the first moment of a variable y (a rather restricted definition of causal effect), that causal effect can be derived from the latent-variable model just as well. The issue is instead merely nonlinearity of the function $E(y|d, x)$.

Angrist also states that the LIM requires "constant coefficient" and "distributional assumptions" for identification,

unlike the LPM. Angrist does not make sufficient qualifications to these assertions, and they are indeed false for the LBM. A random coefficient β in the LBM is permissible without distributional assumption (Ichimura and Thompson 1998) and the coefficients in the standard constant-coefficient LBM are identified semiparametrically (i.e., under unknown distribution of ϵ) under the restriction that d and ϵ are independent, an assumption generally made for causal interpretation of the LPM model as well (Manski 1988; Horowitz 1993; Powell 1994). In addition, as already noted, in a fully saturated model the LPM and the LBM are equivalent in any case.

Is the LATE an Interesting Statistic? The Angrist assertion that LATE is the only interesting statistic if d is endogenous, or perhaps the only statistic worth bothering about because it is the only one identified by the data, has the same partial validity as his preference for the LPM over the LBM. IV is one of the most popular techniques in applied econometrics and has a natural intuition, indeed, one that does not require the LATE interpretation per se. It is one of the most useful tools in the applied economist's kit. The simple IV estimator discussed by Angrist makes minimal assumptions and gives minimal information back to the analyst as a result. But to say that it produces the only statistic of interest does not have defense, both because it is equivalent to an LBM that is fully saturated and may fit the data worse than an LBM if not saturated and because it implies that there is no value to making additional assumptions to obtain additional information.

The limiting nature of the LATE statistic is, again, in its inattention to nonlinearities, interpolation, and extrapolation. That nonlinearities can be important is as true in this case as it was in the exogenous case just discussed when there are additional x covariates and when the model is not saturated. As for interpolation and extrapolation, the LATE statistic $[\bar{y}(z=z_1)-\bar{y}(z=z_0)]/[\bar{d}(z=z_1)-\bar{d}(z=z_0)]$ denotes the effect on a change in z from value z_0 to z_1 on $E(y)$, scaled by the change in the $E(d)$. It does not have implications for the effect on $E(y)$ of any other change in z or for a change in any other policy variable. Thus it is not particularly useful for policy changes other than a change of z_0 to z_1 in the same environment (i.e., conditioned on the same x). This stands in contrast to an LBM for $E(d|x,z)$—or any parametric model for d, for that matter—that allows for the change from z_0 to z_1 to inform policy makers and others of the likely changes of other values of z and of other variables x. In fact, the LATE statistic, which is not a parameter in the usual sense of the word, can always be expressed as a nonlinear combination of parameters of a latent-variable model but not vice versa; hence the latter model is more general than the former. [See Heckman and Vytlacil (1999) for a discussion of how to express the LATE in terms of an LBM; see also Angrist and Imbens (1999) and Heckman (1997, 1999) for a discussion of some of these issues.]

In addition, Angrist's focus on the advantages of the LATE for model-free estimation of the effect of a change in z from z_0 to z_1 on y reveals a philosophical weakness in his position. This is because the effect of a change in z from z_0 to z_1 on y can be estimated from the reduced form; the structural form is not needed. It is not explained why the statistic $[E(y|z=z_1)-E(y|z=z_0)]$ is not the only quantity of interest. The answer that most economists would give is to say that structural coefficients are of interest because they can be used to extrapolate the effects of changes in d on y to other changes in z than from z_0 to z_1 in the same environment; indeed, this is the basic rationale for structural estimation given in the famous essay by Marschak [1953; see also Christ (1994) and Heckman (2000) for discussions; Marschak's argument was more subtle than this, arguing that restrictions on structural parameters are needed for out-of-sample prediction, but this translates into restrictions on the reduced form as well]. However, by ruling out the possibility of learning about any of the effects of changes in z or d other than those induced by a change in z from z_0 to z_1, Angrist removes the need to do structural estimation in the first place. Angrist implicitly assumes that a structural coefficient of interest exists on the variable d in the y equation and that that coefficient has meaning independent of a particular change in z, but the rest of his discussion contravenes that interest.

A reduced-form research program, which is where the Angrist position leads, is of considerable value. There is nothing wrong with a research program to collect a large body of information on the effects of a wide variety of changes in different policy variables (z) from particular values (z_0) to other particular values (z_1), each taking place in particular environments (x). But if the research program stops there, very little useful has been learned other than a collection of facts about the effects of particular policies in particular environments.

Is the Application in the Angrist Article Interesting? The Angrist article uses a recently popular exclusion restriction, or natural experiment, to identify the effect of fertility on labor supply and to illustrate his preferred methods—namely, the use of twins. The use of twins as an exclusion restriction stands in contrast to variations in a government policy or law, which are often used as exclusion restrictions both in recent work on natural experiments as well as a much older literature that uses cross-sectional and overtime variation in state- or country-specific taxes and transfer rules to identify model parameters. However, although it can be of interest to estimate reduced-form policy effects, $[E(y|z=z_1)-E(y|z=z_0)]$, as just discussed, it is not so obvious that there is any interest in estimating the effects of having twins on the expected value of y. Creating twins is not a variable directly subject to policy manipulation.

The difficulty in the use of twins arises for the same reason already discussed—namely, that the preference for model-free estimation of policy effects leads to a lack of interest in the function $E(d|z,x)$ and hence to a lack of interest in what can be learned from a study of twins about the effects of some other, more relevant policy variable that might be manipulated. A model for the $E(d|z,x)$ is needed to make that connection and hence to make a study of twins of any interest, yet that is intentionally eschewed in the approach proposed by Angrist. Once again, this leads to an uninteresting and quite limiting set of exercises.

Conclusions. The set of methods laid out in the Angrist article, at least if a more relevant instrument were used, yields a set of model-free statistics that are suitable for exploratory work on a research question prior to the (possibly nonlinear)

structural estimation and prediction that should be the ultimate object of applied economics research.

ACKNOWLEDGMENTS

This comment is a revision of remarks delivered at the Joint Statistical Meetings, Baltimore, August 1999. I thank Carl Christ, James Heckman, Joel Horowitz, Michael Keane, Thomas Mroz, Geert Ridder, and Edward Vyltacil for comments, with the usual disclaimer that the views expressed here should not be taken as representing those of any of these individuals.

ADDITIONAL REFERENCES

Angrist, J., and Imbens, G. (1999), Comment on "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," by J. J. Heckman, *Journal of Human Resources*, 34, 283–827.

Blundell, R., and Smith, R. (1993), "Simultaneous Microeconometric Models With Censored or Qualitative Dependent Variables," in *Handbook of Statistics* (Vol. 1), eds. G. S. Maddala, C. R. Rao, and H. D. Vinod, Amsterdam: North-Holland, pp. 117–141.

Christ, C. (1994), "The Cowles Commission's Contributions to Econometrics at Chicago, 1939–1955," *Journal of Economic Literature*, 32, 30–59.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462.

——— (1999), "Instrumental Variables: A Reply to Angrist and Imbens," *Journal of Human Resources*, 34, 828–837.

——— (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, 115, 45–97.

Heckman, J., and MaCurdy, T. (1985), "A Simultaneous Equations Linear Probability Model," *International Economic Review*, 18, 21–37.

Heckman, J., and Vytlacil, E. (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

Horowitz, J. (1993), "Semiparametric and Nonparametric Estimation of Quantal Response Models," in *Handbook of Statistics* (Vol. 11), eds. G. S. Maddala, C. R. Rao, and H. D. Vinod, Amsterdam: North-Holland, pp. 45–72.

Ichimura, H., and Thompson, T. (1998), "Maximum Likelihood Estimation of a Binary Choice Model With Random Coefficients of Unknown Distribution," *Journal of Econometrics*, 86, 269–295.

Manski, C. (1988), "Identification of Binary Choice Models," *Journal of the American Statistical Association*, 83, 729–738.

Marschak, J. (1953), "Economic Measurements for Policy and Prediction," in *Studies in Econometric Method*, eds. W. Hood and T. Koopmans, New York: Wiley, pp. 1–26.

Powell, J. (1994), "Estimation of Semiparametric Models," in *Handbook of Econometrics* (Vol. 4), eds. R. Engle and D. McFadden, Amsterdam: North-Holland, pp. 2443–2521.

Comment

John MULLAHY

Departments of Preventive Medicine and Economics, University of Wisconsin, Madison, WI 53705, and National Bureau of Economic Research, Cambridge, MA (*jmullahy@facstaff.wisc.edu*)

Articles bearing titles containing phrases like "Simple Strategies for Empirical Practice" are often wolves masquerading as sheep, the methodologies they devise being far from "simple" and far removed from what most humble practitioners perceive in the realm of "empirical practice." Not so here. Joshua Angrist has written a tight, comprehensive article that is stimulating and important, yet also eminently useful.

As typifies much of Angrist's work, the main concern in this article is on how to elicit interesting characterizations of causal effects from microdata, with the particular twist here being a focus on outcomes measured as "limited dependent variables" (LDV's). The main take-away message I glean from this article is that applied analysts working on causal effect or structural analyses in LDV contexts—traditionally vexing contexts insofar as consistent estimation and inference are concerned—have considerable grounds for optimism. Angrist lays out and interprets systematically a set of issues and methods that provide practitioners with a variety of implementable strategies that might be brought to bear on such empirical problems. A corollary take-away message is that in some respects "this stuff is not really as hard as we've tended to make it," with Angrist demonstrating, for instance, the potential merits of simple linear instrumental variable (IV) methods for estimating causal effects in a variety of LDV contexts. In no event can applied analysts escape the requirement of finding theoretically sound instruments, but Angrist exposits compellingly a broad description of how such instruments might be used to elicit interesting causal inferences in LDV contexts.

I have no real quibbles with any of the substance of Angrist's arguments. Rather, I will devote my commentary mainly to amplifying and expanding several of the themes he develops throughout the article.

1. FOCUS ON CAUSAL OR PARTIAL EFFECTS VERSUS FOCUS ON CONDITIONAL EXPECTATIONS FUNCTIONS

It seems fair to suggest that much of applied microeconometrics is concerned primarily with understanding the signs and magnitudes of quantities like $\partial E[Y|X, D]/\partial(X, D)$ or $\partial E[Y|X]/\partial X$. Yet much of the actual dirty work in undertaking causal analysis in LDV contexts seems to result from decisions to undertake analyses in settings where $E[Y|X, D]$ or $E[Y|X]$ are restricted to be positive without a priori restrictions on parameter values [thus the tradition of using tobit-class conditional expectations functions (CEF's), two-step selection models, exponential CEF's, etc.].

As practitioners, however, we should pause to assess whether specifications akin to $E[Y|X, D] = \exp(X\beta + \alpha D) >$

0 are ultimately buying anything important with respect to the first-order questions being explored. It is only common sense that analysts should spend relatively more energy working on understanding quantities of primary interest (e.g., $\partial E[Y|X, D]/\partial(X, D)$) and worrying relatively less about the formulation and implications of assumptions (e.g., $E[Y|X, D] > 0$) that may have little or no ultimate bearing on the particular questions being addressed. This is especially so when the latter effort (restricting via functional form choice $E[Y|X, D] > 0$) will tend to complicate rather than facilitate the former effort (understanding the partial or causal effects). This is not to suggest that quantities like $E[Y|X]$ and $E[Y|X, D]$ may not themselves be interesting to estimate (e.g., modeling conditional mean health care expenditures for use in forecasting future expenditure levels), but whether $E[Y|X, D]$ is parametrically and parsimoniously better approximated by $\exp(X\beta + \alpha D)$ or by $X\beta + \alpha D$ or by some other $g(X\beta + \alpha D)$ or $h(X, D, \gamma)$ is—even in LDV settings—ultimately an empirical matter to be assessed by goodness of fit (e.g., via conditional moment tests).

I note this important angle somewhat sheepishly since in both articles of mine that Angrist cites (Mullahy 1997, 1998) there is considerable emphasis on the use of exponential CEF specifications as enforcers of positive conditional mean "requirements." In fact, the main motivation for the earlier article was to find a general approach to obtaining consistent estimators of structural parameters in parametric settings less distribution-bound than tobit-class models when the key requirement or side constraint is $E[Y|X, D] > 0$, with the resulting strategy a more or less brute-force method for accomplishing this in a nonlinear IV setting. Angrist's article takes this idea quite a bit further and demonstrates how (and with the addition of a further normalization) the structural parameter estimation approach I discussed can be developed into a model to analyze causal effects when the latter are characterized as "proportional treatment effects."

Whether a proportional, as opposed to an additive, treatment effect is interesting (it may or may not be) remains to be seen in any particular application. Wooldridge's recent work on average causal effects (ACE's; Wooldridge 1999) would seem pertinent in contexts in which the role of the X covariates is more "intrusive" than in the simple additive/linear treatment-effect setup. As a general matter, the extent to which inferences about causal effects—however formulated—hinge on the inclusion of particular covariates as conditioning regressors is an interesting and potentially important angle that Angrist's article begins to explore and is one that likely to merit additional research. For the particular issues at hand, the very nature of proportional treatment effects brings into focus explicitly the role of the X's, so how their correlation with the treatment indicators (D) influences inferences about the causal impacts of the latter would seem to be a first-order consideration.

2. TWO-PART MODELS

I found Angrist's analysis of the interpretation of causal effects in two-part models to be extremely illuminating. In a reduced-form setting, I argued (Mullahy 1998) that, amid all the concerns about zeros, robustness, outliers, transformation, retransformation, and such that have typically attended modeling efforts involving two-part models, it would be surprising in a well-structured empirical investigation if there were not stated or lurking concerns about partial effects $\partial E[Y|X]/\partial X$ and/or CEF's $E[Y|X]$ themselves. It may be in some cases that such concerns are not articulated, and it may even be in less well-structured problems that they are not obvious to analysts themselves. Yet, for instance, unless robustly estimated β's from a log-linear part-2 specification of a two-part model could inform a first-order question about a partial effect or a CEF, of what practical use were they likely to be? My rather modest and rather obvious argument was just that the concerns about the $\partial E[Y|X]/\partial X$ or the $E[Y|X]$—if they are the reason the analysis is being conducted in the first place—should enjoy first-order prominence in the estimation exercise and that concerns about zeros, outliers, and such should be relegated in some sense to second-order status. In some cases, two-part models will serve nicely to address such first-order concerns—at least in reduced-form settings—but in others they will not.

Angrist appears to have some sympathy with such arguments, but more importantly he provides a valuable service to users of two-part models by unearthing one major limitation in the analysis of causal effects. The problem is not with part 1 (i.e., the logit, probit, or linear probability component) because part 1 of the two-part model falls within the main lines of Angrist's LDV analysis. Rather, the complication arises with part 2 of the two-part model in which the additional "$Y > 0$" conditioning arises. In the standard (reduced-form) two-part model, quantities like $E[Y|X, Y > 0]$ are prominent and identified readily. But in the counterfactual settings in which causal effects are manifested, wherein the realized Y arises from self-selection into or out of treatment, such selection effects introduce an ambiguity [Eq. (12)] and thereby confound the ability to glean causal effects from part-2 estimates (unless, as Angrist notes, censored regression methods are used, but Angrist also offers some compelling arguments against blind reliance on censored regression approaches). As such, if estimation of and inference about causal effects are the primary analytical concerns in data settings with $Y \geq 0$ and $\Pr(Y = 0) > 0$, analysts may be well advised to avoid two-part modeling strategies and pursue some of the more direct linear and nonlinear estimation approaches discussed by Angrist in which one-part estimation approaches—zeros and all!—yield estimates of causal effects that will be directly interpretable.

3. BEYOND CEF'S

One noteworthy feature of the methods Angrist discusses—with the main results attributed to Abadie (1999)—are their applicability to estimation of causal effects for characteristics of the conditional distribution $\phi(Y|X, D)$ beyond just CEF's. Estimation of causal or treatment effects for conditional quantiles (QTE's) and conditional distribution function ordinates ("distiles"), $\Pr(Y < c|X, D)$, is demonstrated to fit properly within the weighting strategies advanced by Abadie. Since

quantile and distile analyses may be more relevant in addressing particular analytical questions or policy issues than CEF's, the ability to extend the mechanisms of causal-effect estimation strategies in such directions should be welcomed by applied researchers.

I might suggest that Angrist's claim about an advantage of quantile causal analysis over distile causal analysis, being that the latter relies on application-specific values of ordinates (c), is perhaps oversold or underdeveloped. In many interesting and policy-relevant applications, focus on the application-specific ordinates is fundamental: What factors cause employers to offer more than one health insurance plan to employees? What factors cause consumers of alcoholic beverages to drink more than two drinks per day? What factors cause elderly individuals to utilize more than one prescription drug product? What factors cause second earners to work more than 17 hours per week? Each of these concerns an important real-world causal question for which the application-specific ordinate—whether for institutional, legal, physiological, or other reasons—is a fundamental "pivot point" for the analysis. Analysis of such distile relationships—whether in causal or in reduced-form settings—is a potentially powerful method for informing decision makers, and it seems to me that its merits relative to quantile analyses would have to be judged on a case-by-case basis.

4. CONCLUSIONS

Practitioners of applied microeconometrics will find that the time invested in a careful reading of Angrist's article has sizable returns, not merely because of the analytical and conceptual insights it offers but ultimately also because it suggests a variety of important practical innovations that applied researchers can exploit empirically to more clearly understand the nature of causal relationships in their data. Applied micro researchers in areas like health, labor, development, public finance, and such work commonly with data wherein (a) outcome measures are limited—often most importantly because they are nonnegative with mass points at 0—and (b) the roles of conditioning covariates (X's) are most conveniently summarized via regression methods. Angrist's article provides applied researchers so endowed with a set of powerful tools for eliciting causal inferences from such data. Of course, the punch line is familiar and inevitable: All bets are off without theoretically solid instruments. But with such instruments in hand, the strategies exposited by Angrist offer analysts a rich set of perspectives on estimation.

ADDITIONAL REFERENCE

Wooldridge, J. M. (1999), "Estimating Average Partial Effects Under Conditional Moment Independence Assumptions," mimeo, Michigan State University, Dept. of Economics.

Comment

Petra TODD
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104 (*petra@athona.sas.upenn.edu*)

This article considers ways of estimating the effect of binary, endogenous regressors in models with limited dependent variables. It questions the usefulness of conventional estimation strategies aimed at recovering structural model parameters and advocates the use of simple instrumental variables (IV) estimators as an alternative, on the grounds that these estimators invoke weaker assumptions and often suffice to answer questions of interest in empirical studies. In the author's view, the main challenge facing empirical researchers is the problem of identification of "treatment effects" through IV. Here, I expand the discussion by considering two questions: (1) What are the limitations of the causal model as a paradigm for policy analysis? (2) When simple estimators are suitable for answering a question of interest, what are the trade-offs that need to be considered in using them?

1. WHEN IS THE "CAUSAL MODEL" A USEFUL PARADIGM FOR POLICY ANALYSIS?

The article gives the impression that most interesting questions in economics can be answered within the context of the "causal model" [variously attributed to Neyman (1923), Fisher (1935), Cox (1958), Roy (1951), and Rubin (1978)]. The causal model is a very general framework that assumes that there are potential outcome states, associated with having received some treatment and having received no treatment (or possibly a different treatment). For each individual, only one of the potential outcome states is observed, which leads to a missing-data problem in attempting to draw inferences about aspects of the treatment effect distribution.

The language of potential outcomes is very general, so almost any economic problem could be formulated in these terms. However, this does not mean that estimators proposed in the literature for the causal model are useful for all or even most economic problems. A major limitation of the model is that it assumes that the state "with treatment" has been observed for at least a subset of people. In medical trials or biological experiments, this assumption is probably reasonable, but in economics we often are interested in evaluating effects of treatments that have not yet been implemented. For example, we might be interested in predicting the effect of raising the age receiving Social Security benefits to a new age or the effect of introducing new term limits on welfare participation. An implicit assumption needed to apply any of the identification results described in this article is that both Y_1

and Y_0 are observed, which would not be satisfied in many cases. Thus, the causal model and the estimators developed for it are inadequate when it comes to any economic question involving a state of the world that has not been observed.

Despite these limitations, treatment effect estimates are sometimes used to predict consequences of new treatments or of existing treatments for new populations. When the policy change being considered is very close to changes observed in the data and is to be applied to similar individuals, then extrapolation may be reasonable. But if the policy change is (a) of a magnitude outside the range of experience, (b) along a new dimension, or (c) intended for a new population, then it is an open question as to how reliable the estimates would be. Treatment impact estimates are not intended to capture anything invariant to changing conditions. The assumptions that would be required to justify extrapolations or generalizations are often quite strong. To some extent, it is an illusion that these estimators allow for general kinds of inference under weak assumptions because the assumptions required to justify their application for particular purposes are often not made explicit.

Angrist argues that the identification results for the local average treatment effect (LATE) estimator provide a "foundation for credible causal inference" (p. 5) and a "minimum controversy jumping-off point for any prediction exercise" (p. 5). But as discussed previously, the causal model maintains such a high level of generality that it offers little guidance for many kinds of problems. It is important to recognize when and when not the LATE estimator is likely to be useful. This article claims that "in practice, estimates of LATE differ little from estimates based on the stronger assumptions invoked to identify effects on the entire treated population" (p. 5). But such an extrapolation is not justified under the theory and is surely unlikely to hold in all circumstances. As shown in a number of works by various authors, LATE provides the average effect of treatment for the group of "compliers," who are people induced by the instrument to receive treatment. This group does not, except under very special circumstances, correspond to the group of treated persons and may, in fact, represent only a small fraction of persons receiving treatment. So, not only does LATE not suffice for answering problems of the sort described by (a), (b), and (c), it also does not generally provide the average effect of treatment on the treated—the most common parameter of interest in analyzing social programs. Another not-so-attractive feature of the estimator is that the population of "compliers" is usually not identifiable, so we cannot say exactly whose treatment impact is being estimated. Applying the LATE parameter estimate to the full population of treated persons would generally require ruling out certain types of heterogeneity in individual responses to treatment, as discussed by Heckman (1997). The explicit assumptions that could justify this type of extrapolation were shown by Heckman and Vytlacil (2000, in press).

2. ESTIMATORS SHOULD BE TAILORED TO THEIR APPLICATION

A general theme throughout the article is that whenever possible researchers should restrain from imposing parametric restrictions in estimation because the structure they impose may be wrong, resulting in biased estimates. This view is expressed, for example, in the discussion of the two-part model. The article advocates the use of a two-part approach instead of a conventional selection-model approach, maintaining that the two-part approach is less likely to lead to biased parameter estimates because it does not impose cross-equation restrictions as a conventional tobit estimator would.

How do researchers decide whether or not to impose restrictions across estimating equations that are structurally linked? The most agnostic approach would be to estimate both equations completely nonparametrically, but this approach is rarely practical. Nonparametric estimators are consistent under the most general conditions, but in conventional size samples the standard errors would be so large that the estimates would for practical purposes be useless. Researchers impose structure because they are willing to restrict the class of models under which their estimators are consistent in exchange for greater precision. These efficiency considerations would also apply in the deciding whether or not to impose cross-equation restrictions in estimating simultaneous-equations models. If we are right about the structure, there is an efficiency gain from imposing the restrictions. If we are wrong, imposing them may lead to bias. When there are overidentifying restrictions, it is at least possible to develop tests of the model specification that can serve as a guide in choosing a structural model that is supported by the data. This article presents a dichotomy between the two-part model and a fully parametric, traditional selection specification, but there are many other modeling choices, ranging from fully nonparametric to semiparametric to fully parametric. In modest size samples, efficiency is an important consideration and a parametric model may be the most appropriate one. For very large samples, fewer parametric approaches are feasible. Both the probability-of-working and the hours-worked equations could, for example, be estimated by a semiparametric method, such as semiparametric least squares (Ichimura 1993). Which estimation method is most appropriate ultimately depends on how much data is available and how many parameters are to be estimated. It does not make sense to criticize the use of structured approaches in favor of more flexible modeling approaches without regard to the context in which they are being applied.

3. ON THE GAP BETWEEN THEORY AND EMPIRICAL PRACTICE

Finally, another recurring emphasis in the article is on the value of shortcuts in empirical practice. In the author's view, there is a gap between the econometric theory and what is feasible in practice, and shortcuts and approximations help to bridge this gap. To this end, the article extols, for example, the virtues of ordinary least squares (OLS) as a way of approximating any unknown conditional expectation function, without much concern as to whether a best *linear* approximation is a *good* approximation. If the model has discrete regressors and is fully saturated, it is well known that OLS is equivalent to a nonparametric estimator. But in other cases, a linear approximation could be highly biased and the bias will not go away as the sample size gets large. There are a variety of alternative asymptotically unbiased estimators. Along similar lines,

the article suggests that, in implementing a "selection-on-observables" estimator, it may be all right simply to ignore the weighting scheme. Such claims seem to me fairly outrageous because failure to take into account weighting could lead to very different estimates and would not provide a consistent estimator of the parameter of interest. Heckman, Ichimura, and Todd (1998) discussed alternative estimators that are not difficult to implement and take into account the weighting scheme.

Finally, the appendix presents a variety of shortcuts for obtaining estimates and standard errors—some of which lead to inconsistent estimates. When shortcuts lead to very crude approximations or inconsistent estimates, I question their usefulness. I do think it unreasonable to expect that estimates based on simple IV estimators be accompanied by correct standard errors that correctly take into account the influence of estimated parameters of estimated regressors. I see little value in reporting incorrect standard-errors estimates on the grounds that it saves time in computation, especially when there are several works long available in the literature that develop simple ways of solving these kinds of problems (e.g., see Newey 1984). Although there is surely a gap between the estimators developed in the theoretic econometric literature and the estimators used in the empirical literature, the gap is not as great as this article supposes.

ADDITIONAL REFERENCES

Cox, D. R. (1958), *The Planning of Experiments*, New York: Wiley.

Fisher, R. A. (1935), *Design of Experiments*, London: Oliver and Boyd.

Heckman, J. J. (1997), "Instrumental Variables: Evidence From Evaluating a Job Training Programme," *Journal of Human Resources,* 32, 441–462.

Heckman, J., Ichimura, H., and Todd, P. (1997), "Matching as an Econometric Evaluation Estimator: Theory and Evidence on its Performance Applied to the JTPA Program, Part I. Theory and Methods," *Review of Economic Studies,* 64, 605–654.

——— (1998), "Matching as an Econometric Evaluation," *Review of Economic Studies,* 65, 261–294.

Heckman, J., and Vytlacil, E. (2000), "Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluations of Social Programs," unpublished manuscript, University of Chicago, Dept. of Economics.

——— (in press), "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, eds. C. Hsiao, K. Morimune, and J. Powell, Cambridge, U.K.: Cambridge University Press.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics,* 58, 71–120.

Neyman, J. (1923), "Statistical Problems in Agricultural Experiments," Supplement to the *Journal of the Royal Statistical Society,* 2, 107–180.

Newey, W. (1984), "A Method of Moments Interpretation of Sequential Estimators," *Economic Letters*, 14, 201–206.

Roy, A. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers,* 3, 135–146.

Rubin, D. (1978), "Baysian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics,* 6, 34–58.

Reply

Joshua D. Angrist

I thank the discussants and the session organizer, outgoing *JBES* editor Jeff Wooldridge, who will be missed by associate editors and *JBES* authors alike. The comments in this session offer a variety of viewpoints, some sympathetic, others more critical. I am pleased to be part of this stimulating and constructive exchange and look forward to more interaction of this type in the future.

Hahn. My article starts by suggesting that a focus on treatment effects simplifies the problem of causal inference in limited dependent variable (LDV) models with endogenous regressors. Endogeneity's half-sister is the assumption of unobserved individual effects, since the omitted-variables motivation for fixed effects in panel data is similar. Hahn asks whether a shift in focus from index parameters to average causal effects might also simplify the notoriously difficult problem of working with nonlinear models for binary panel data (e.g., as in Card and Sullivan 1988). Hahn shows that here too the causal-effects framework pays off, demonstrating that average causal effects are easily estimated in a fixed-effects probit model with a binary regressor even though the probit coefficient is not. A second noteworthy feature of Hahn's comment is his observation that identification in this framework turns on careful modeling of the relationship between the assignment variable and the fixed effects. Traditional econometric models are primarily concerned with the stochastic process generating outcomes. A shift of attention toward modeling the assignment mechanism is consistent with the notion that observational studies should try to mimic the sort of controlled comparisons generated by experiments.

Imbens. Imbens begins by endorsing the view that the substantive goals of empirical research are better served by the potential outcomes/causal framework than by structural modeling. Going one step further, Imbens notes that the problems with "model-first econometrics" are not limited to LDV's. In the same spirit, he sensibly wonders (as did Wooldridge 1992) why so much attention has been devoted to estimating the Box–Cox transformation model, which does not actually identify $E[Y|X]$ without a distributional assumption. Imbens also nicely articulates the rationale for looking at the twins example and notes that this is similar to the rationale for extrapolating from clinical trials in narrow subpopulations. And, in fact, I believe the "twins experiment" is likely to have predictive value in a variety of situations, though any extrapolation will probably be improved by control for demographic characteristics.

In spite of (or perhaps because of!) our years of discussion and collaboration, Imbens and I still do not agree on certain things. He feels strongly that inference with discrete outcomes should use models and procedures that respect inherent non-

linearities. The HIRZ procedure discussed in his comment is an intriguing way to do this, and my article discusses others. I remain unconvinced, however, that nonlinearity is of major substantive importance in empirical work with binary regressors. The proof of the pudding, I suppose, would be to show that HIRZ or some of the methods I discuss in the article lead to better predictions or decisions than the linear workhorse.

Moffitt. Moffit provides a valuable counterpoint. He begins by arguing that latent-index models have more predictive power than reduced-form causal models. Such claims seem hard to establish a priori, though I do not doubt that in some cases this may be true. But empirical strategies in which the bulk of the research effort goes into estimating index coefficients may fail to answer the most basic question in empirical work: What happened here? Causal models and local average treatment effect (LATE) answer this question, while index coefficients alone are not sufficient for causal inference. I agree, however, that structural modeling can be useful for going beyond basic questions, and I have used index and other structural models for this purpose. Moffit is right, however, in suggesting that I find the model-free analysis he refers to as "exploratory" more interesting than structural extrapolation. This seems to me to be where the potential for real discovery lies.

Moffitt asks tough but fair questions about the twins application. He is correct to point out that twinning is not a manipulable policy instrument. That is one reason I do not limit the analysis to reduced-form effects of multiple births. Yet, to my mind, the best evidence we have on the controversial question of the consequences of out-of-wedlock childbearing comes from the twins experiment (Bronars and Grogger 1994). And the economic consequences of childbearing are of interest in many other contexts. As noted in the introduction to my article, the twins experiment can help determine whether fertility reductions are responsible for trend increases in female labor supply. Beyond this sort of question, the twins results seem useful for assessing the economic consequences of increased abortion access in the United States, the dissemination of contraceptive implants in developing countries (e.g., see Kane, Farr, and Janowitz 1990), and various programs to encourage or require implant use by welfare mothers (Haveman and Wolfe 1993).

Finally, it is fair to question the external validity of twins estimates, as Moffit does toward the end of his comment. Angrist and Evans (1998) showed that the labor-supply consequences of twinning can be reconciled with estimates of the labor-supply consequences of childbearing resulting from a very different natural experiment induced by preferences for mixed-sex sibling pairs. This suggests that results from both experiments may be quite robust and may even be "structural" in the sense that they tell us something general about the economic response to childbearing.

Mullahy. Mullahy brings a welcome health economist's perspective to the discussion. Health economists have long grappled with the issues discussed in my article and, as I hope I made clear, I drew ideas and inspiration from Mullahy's work on models with nonnegative dependent variables. Mullahy's comment further highlights exactly the sort of utilitarian trade-off that underlies the choice of empirical strategy. On one hand, we would like our models to be consistent with any functional form restrictions known to be true. On the other, attempts at this may complicate estimation and inference, for little payoff. Worst of all, the technical complexity of nonlinear models seems to cause authors to "wax econometric," an effort that may come at the expense of attention to substantive issues of importance (an example here is the long series of papers discussing exactly how data from the RAND health-insurance experiment should be analyzed). Mullahy's heretical discussions of "one-part models," here and in earlier work, represent a refreshing effort to keep the empirical train on track. Mullahy's point that application-specific distribution ordinates can provide more useful summary information than quantiles is also clearly worth thinking about.

Todd. Todd provides a useful and clear overview of the causal framework, which she suggests is better suited for medical questions than economic questions. The thrust of her argument is that there is too much heterogeneity in economic interventions and outcomes for the causal approach to be very useful for economists. But my perception is that medical research is even more explicitly concerned with heterogeneity than economics. This is evidenced by the many qualifications regarding patient characteristics, conditions, and interactions in medical treatment protocols. A difference between the prevailing medical research paradigm and the traditional econometric perspective, however, is whether the question of heterogeneous response is to be addressed by better theory or more evidence. Of course, theory helps us decide where to look for heterogeneity, but as Rosenbaum (1999, p. 301) put it: "If a treatment has substantially different effects in different types of people, then we need to *discover* this (italics mine). Rosenbaum also noted that "randomized clinical trials routinely look for heterogeneous treatment effects, and they do this without representative samples from national populations."

On the technical side, Todd suggests I advocated use of the two-part model because it flexibly approximates nonlinear functional forms for nonnegative outcomes. Although I noted the flexibility, the upshot of my discussion was that the two-part model is poorly suited to causal inference because of the difficulty of interpreting part 2. Finally, Todd points out the failure to allow for sampling variance in the scaling factors used to generate effects for nonlinear models. This is indeed a hard-to-defend shortcut, though I shall take a stab at defense by noting that randomness of the scaling factor, which typically estimates a sample average, is likely to be a minor source of sampling uncertainty.

ADDITIONAL REFERENCES

Card, D. E., and Sullivan, D. (1988), "Effects of Subsidized Training on Movements In and Out of Employment," *Econometrica*, 56, 497–530.

Haveman, R., and Wolfe, B. (1993), "Children's Prospects and Children's Policy," *Journal of Economic Perspectives*, 7, 153–174.

Kane, T., Farr, G., and Janowitz, B. (1990), "Initial Acceptability of Contraceptive Implants in Four Developing Countries," *International Family Planning Perspectives*, 16, 49–54.

Rosenbaum, P. R. (1999), "Choice as an Alternative to Control in Observational Studies," *Statistical Science*, 14, 259–304.

Wooldridge, J. (1992), "Some Alternatives to the Box–Cox Regression Model," *International Economic Review*, 33, 935–955.