TECHNICAL WORKING PAPER SERIES

ESTIMATION OF LIMITED-DEPENDENT VARIABLE MODELS WITH DUMMY ENDOGENOUS REGRESSORS: SIMPLE STRATEGIES FOR EMPIRICAL PRACTICE

Joshua D. Angrist

Technical Working Paper 248
http://www.nber.org/papers/T0248

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2000

This paper was prepared for the 1999 ASA meetings in Baltimore, Maryland. I thank session discussants, Jinyong Hahn, Robert Moffitt, Petra Todd, and Jeff Wooldridge for their comments. Thanks also go to Daron Acemoglu, Bill Evans, Bill Greene, Guido Imbens, John Mullahy, and Steve Pischke for helpful discussions and comments on an earlier draft, to Melissa Schettini for outstanding research assistance, and especially to Alberto Abadie who provided extensive helpful comments and shared his computer programs. Data and computer programs used here are available on request. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

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

Joshua D. Angrist

NBER Technical Working Paper No. 248
January 2000
JEL No. C34, C35, J22

ABSTRACT

Applied economists have long struggled with the question of how to accommodate binary endogenous regressors in models with binary and non-negative 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, a number of simple strategies is 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 paper are illustrated by using multiple births to estimate the effect of childbearing on employment status and hours of work.

Joshua D. Angrist
MIT Department of Economics, E52-353
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
angrist@mit.edu

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, where the 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, since they involve binary regressors. 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 principle challenge facing empirical researchers conducting studies of this type is identification. Successful identification in this context usually means finding an instrumental variable 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 non-negative with a mass point at

zero. Examples of this sort of outcome include earnings, hours worked, and expenditure on health care. The analysis of such limited dependent variables (LDVs) 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 paper argues that the difficulty with endogenous variables in nonlinear LDV models is, in fact, more apparent than real. For binary endogenous regressors, at least, the technical challenges posed by limited-dependent variable 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 LDVs is not fundamentally different from causal inference with continuous outcomes.

The next section 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 the identification is for 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 (at least to me). Moreover, claims for identification in this context typically strike me as overly ambitious, since even an ideal randomized experiment fails to identify conditional-on-positive 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 two-stage least squares (2SLS) are no less appropriate for LDVs 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 semi-parametric estimators. The first, due to Mullahy (1997), is based on a multiplicative model that can be estimated using a simple nonlinear instrumental variables (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 earlier semi-parametric 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" has been 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 What to estimate?

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 marital fertility have been concentrated in the 2-3 child range.

What is the object of study 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, at least, 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 earnings, 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 exogenous declines in fertility have been a causal factor in increasing female employment rates over time. In contrast with descriptive analyses, causal relationships answer counter-factual questions, and are therefore more likely to be of value for predicting the effects of changing policies or changing circumstances (see, e.g., Manski [1996]).

Causal relationships can be described most simply using explicit notation for counterfactuals or *potential outcomes*. This approach to causal inference was developed by 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 child-bearing 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 hand side is observed, but the second term is an unobserved counter-factual average that we assume is well-defined.

The right hand 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 other 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:

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

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, since childbearing is affected by choices 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 determining 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 ν_{wi},ν_{mi}), as in Heckman (1974) and many other papers:

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

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

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 behavior 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:

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

where

$$Y_{0i} = 1(X_i'\delta > -\varepsilon_i)(X_i'\delta + \varepsilon_i) \tag{5a}$$

$$Y_{1i} = 1(X_i'\delta+\varphi > -\varepsilon_i)(X_i'\delta + \varphi + \varepsilon_i). \tag{5b}$$

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:

$$E[Y_{1i}-Y_{0i}|\, D_i=1] = E\{1(X_i'\delta + \varphi > -\varepsilon_i)(X_i'\delta + \varphi + \varepsilon_i)-1(X_i'\delta > -\varepsilon_i)(X_i'\delta + \varepsilon_i)|\, X_i'\gamma > \eta_i\}. \tag{6}$$

Note, however, that knowledge of the parameters on the right hand side of (6) is still not enough to evaluate this expression. The following Lemma outlines the identification possibilities in this context:

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

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

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

iii. (Imbens and Angrist, 1994) 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.

iv. (Angrist and Imbens, 1991) Suppose that LATE is identified as in (iii), 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 $E[Y_{1i}-Y_{0i}|\, D_i=0]$.

This set of results can easily be summarized using non-technical 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, i.e., for exogenous treatment, the LATE result, and the specialization of LATE to effects on the treated and non-treated, require no assumptions beyond those in the lemma. In fact, no information of any kind about the structural model is required for identification in these cases. On the other hand, the result for endogenous treatments in part (iii) does not 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 effects mentioned in part (iii) capture 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, call 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 since 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. But numerous examples (including the one below) lead me to believe that even though valid instruments guarantee identification only for causal effects on compliers, in practice, estimates of LATE differ little from estimates based on the stronger assumptions invoked to identify effects on the entire treated population.

To sum up: I see parameters like LATE and the effect of treatment on the treated as providing a minimum-controversy jumping-off point for causal inference and prediction, whether or not the dependent variable is limited. For some readers, however, a focus on causal effects of any type may seem misguided. After all, it is structural parameters that are usually linked to economic theory. But what was the purpose of the theory in the first place? The ultimate goal of theory-motivated structural estimation seems to differ

little from the agenda outlined here. For example, Keane and Wolpin (1997, p. 111) want to use 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 natural pre-requisite for this is credible assessment of the consequences of past changes. Structural parameters that are not linked to causal effects are not useful for this basic purpose (Mullahy [1998] makes a similar point).

2. Causal effects on LDVs

2.1 Average effects in experimental data

In this section, I return to the non-structural causal model defined directly using potential outcomes, and ask whether the limited nature of the dependent variable has any implications for empirical analysis. A natural starting point for this discussion is the analysis of randomized experiments, since some of the issues raised by LDVs 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:

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

If D_i is also independent of Y_i, 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 or not Y_i is binary, non-negative, or continuously distributed. If Y_i is binary, then the difference in means on the left hand side of (8) estimates a difference in probabilities, while if Y_i has a mass point at zero, the difference in means estimates 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 non-negative 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, first noted in this context by McDonald and Moffit (1980):

$$E[Y_i| D_i=1] - E[Y_i| D_i=0] = \{P[Y_i>0| D_i=1] - P[Y_i>0| D_i=0]\}E[Y_i|Y_i>0,D_i=1] + \{E[Y_i|Y_i>0,D_i=1] - E[Y_i|Y_i>0,D_i=0]\}P[Y_i>0| D_i=0]. \quad (9)$$

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, see Evans, Farrelly, and Montgomery (1999), who analyze 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? A large literature contrasting two-part and sample-selection models for LDVs suggests this issue remains controversial. (See, for example, Duan, et al, 1984; Hay and Olsen, 1984; Hay, Leu, and Rohrer, 1987; Leung and Yu, 1996; Maddala, 1985; Manning, Duan, and Rogers, 1987; and 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):

$$E[Y_i| Y_i>0, D_i=1] - E[Y_i| Y_i>0, D_i=0] = E[Y_{1i}| Y_{1i}>0, D_i=1] - E[Y_{0i}| Y_{0i}>0, D_i=0]$$
$$= E[Y_{1i}| Y_{1i}>0, D_i=1] - E[Y_{0i}| Y_{0i}>0, D_i=1] \quad (10a)$$

$$= E[Y_{1i}-Y_{0i}|\ Y_{1i}>0,\ D_i=1]$$

$$+ \{E[Y_{0i}|\ Y_{1i}>0,\ D_i=1] - E[Y_{0i}|\ Y_{0i}>0,\ D_i=1]\}. \qquad (10b)$$

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 would not tell is 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].

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

$$Y_i = 1[Y_i^*>0]Y_i^*, \text{ where} \qquad (11)$$

$$Y_i^* = Y_{0i}^* + (Y_{1i}^*-Y_{0i}^*)D_i = Y_{0i}^* + D_i\alpha.$$

Two recent studies with Tobit-type censoring in a female labor supply models are Blundell and Smith (1989) and Lee (1995), both of which include endogenous 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. What sort of causal parameter is α? One answer is that α is the causal effect of D_i on Y_i^*, but 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 causal effect on conditional-on-positive Y_i (for details, see the appendix). In particular,

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

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

Thus, the censored-regression model does succeed in separating causal effects from selection effects in conditional-on-positive comparisons. (We don't know 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. While I might comfortably use a censored regression model to analyze CPS earnings data, since these data are in fact censored at the CPS topcode, the notion of a latent labor supply equation that can take on negative values is less clear cut. The mass point in this case comes about because some people choose to work zero hours, and not because of a problem in the measurement of outcomes (Maddala [1985] makes a similar point for 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 equates wages to marginal rates of substitution. The latent structure in this case cannot be interpreted without abstract theoretical constructs, and, most importantly, the estimated index coefficients have no predictive value for directly observable quantities.

Second, even if we adopt a theoretical framework that makes conditional-on-positive effects meaningful, identification of the censored-regression model requires assumptions beyond those needed for identification of unconditional effects. Semi-parametric 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 the distributional assumptions required for identification, the causal interpretation of α in (12) turns heavily on the additive, constant effects model in (11). This is because objects like $E[Y_{1i}-Y_{0i}|\ Y_{1i}>0]$ involve the *joint* distribution of Y_1 and Y_0. The constant-effects assumption nails this joint distribution down, but the actual data contain information on marginal

distributions only (which is why even randomized trials fail to answer the causal question that motivates sample selection models).

These concerns, echoed by many applied researchers, have stimulated a search for alternative strategies for the analysis of non-negative outcomes (see, e.g., Moffit's [1999] recent survey). The two-part model (2PM), introduced by Cragg (1971), and the subject of recent discussion among health economists, seems to provide a less demanding framework for the analysis of LDVs 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 freely 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 (see, e.g., Eichner, *et al*, 1997, for a recent application to expenditure data). Log-linearity for the second part may be desirable since this imposes non-negativity of fitted values.

One attraction of the 2PM is that it fits a nonlinear functional form to the conditional expectation function (CEF) for LDVs even if both parts are linear. On the other hand, 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 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 $\varphi(\bullet)$ and $\Phi(\bullet)$ are the standard Normal density and distribution functions and σ is the standard deviation of ε_i (see, e.g, McDonald and Moffit, 1980).

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 would therefore seem to be of limited interest for the type of estimation problems that are the primary focus of this paper.

2.3 Effects on distributions

Interest in conditional-on-positive effects and sample selection models sometimes reflects an interest in the consequences of D_i beyond the impact on average outcomes. A natural question is whether there are schemes for estimating effects on distributions that are less demanding than sample selection models. I believe the answer to this question is "yes" since, 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 and 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 changes the probability $Y_i=0$, as in the first part of the 2PM, or whether there is a change in the distribution of outcomes at any positive value, or over any interval of positive values. This information is enough to make social welfare comparisons, as long as the comparisons of interest do not involve the joint distribution of Y_{1i} and Y_{0i}.

2.4 Covariates and nonlinearity

The conditional expectation of Y_i given D_i is linear, as are other conditional relationships involving only D_i. Suppose, however, that identification is based on a "selection-on-observables" assumption instead of presumed random assignment. This means causal inference is based on the presumption that

$Y_{0i} \perp\!\!\!\perp D_i | X_i$.

Now causal effects must be estimated after conditioning on X_i. For example, the effect of treatment on the treated can be expressed as:

$$E[Y_{1i} - Y_{0i} | D_i=1] = E\{E[Y_{1i} | X_i, D_i=1] - E[Y_{0i} | X_i, D_i=1] | D_i=1\} \quad (14)$$
$$= \int \{E[Y_{1i} | X_i, D_i=1] - E[Y_{0i} | X_i, D_i=0]\} P(X_i=x | D=1) dx.$$

Estimation using the sample analog of (14) is straightforward if X_i has discrete support with many observations per cell (see, e.g., Angrist, 1998). Otherwise, some sort of smoothing (modeling) is required to estimate the possibly nonlinear CEFs $E[Y_{1i}|X_i, D_i=1]$ and $E[Y_{0i} | X_i, D_i=0]$.

Often, 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 $E[Y_i|X_i, D_i]$ (see, e.g., 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 $E[Y_{0i} | X_i, D_i=0]$, though this leaves the problem of estimating $P(X_i=x | D=1)$ to compute the average difference in CEFs. 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 the empirical treatment-effect heterogeneity may be limited enough that different 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 estimating using distributional assumptions and maximum likelihood (ML). An early and influential paper in this mold is Heckman (1978). This approach is not wedded to ML; Heckman (1978), Amemiya (1978, 1979), Newey (1987), and Blundell and Smith (1989) discuss two-step procedures, minimum-distance estimators, GLS estimators, and other variations on this framework. Semi-parametric estimators based on weaker distributional assumptions are 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 variable, Z_i. So the complete model is:

$$Y_i = 1[Y_i^*>0]Y_i^*,$$

$$Y_i^* = Y_{0i}^* + (Y_{1i}^*-Y_{0i}^*)D_i = Y_{0i}^* + D_i\alpha = \mu + D_i\alpha + \varepsilon_i$$

$$D_i = 1[\gamma_0+\gamma_1 Z_i > \eta_i]. \tag{15}$$

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. Semi-parametric estimators typically work by substituting an estimated conditional expectation, $\hat{E}[D_i| Z_i]$, for D_i and then using a non-parametric or semi-parametric procedure to estimate the pseudo reduced form (e.g., Manski's [1975] maximum score estimator for binary outcomes).

In the previous section, I listed problems with this approach: 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 develop semi-parametric method by Blundell and Powell [1999] for continuous endogenous variables). Second, this

approach turns heavily on the censored-regression/latent-index/constant-coefficients setup. We can add to these two 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 (see, e.g., Lee, 1996).

A final point from my perspective 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 [iv] in the Lemma). It is almost certainly of interest to extrapolate from this group's experiences to those of other women, but that problem is distinct from identifying the causal effect of childbearing in the "twins experiment".

4. New econometric methods

Although conditional moments and other probability statements involving only D_i are linear, 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 an implicit concern that, since the covariate parameterization is not saturated, LDVs lead to nonlinear CEFs. The 2PM is also sometimes motivated this way (see, e.g., Duan, *et. al,* 1984).

Are there any 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 described the relationship of interest:

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

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

These assumptions lead to the formulation,

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

an equation that is easily estimated by 2SLS. (If we substitute "population-regression" for "conditional expectation" in [17a], the instrument must be independent of the regression residual instead of independent of Y_{0i} given X_i).

Because D_i is binary, it is tempting to use a nonlinear model such as Probit or Logit for the first stage for this 2SLS problem. In the context of an additive constant-effects model such as (17), however, second-stage estimates computed by OLS regression on first-stage fitted values from a nonlinear model 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. Alternately, 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 since $P[D_i=1| X_i, Z_i=1]=1$ 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 other non-negative LDVs. This motivates Mullahy (1997) to estimate causal effects on non-negative LDVs 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. 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 and the instrument are assumed to be independent of potential outcomes:

$$Y_{0i} \perp\!\!\!\perp (D_i, Z_i) \mid X_i, \omega_i \tag{M1}$$

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

$$Z_i \perp\!\!\!\perp \omega_i \mid X_i, \tag{M2}$$

though ω_i and D_i are presumed to be related. The CEF for Y_{0i} is constrained to be non-negative using an exponential model and (M1):

$$E[Y_{0i} \mid D_i, Z_i, X_i, \omega_i] = exp(X_i'\beta + \pi\omega_i) = \omega_i^* exp(X_i'\beta), \tag{M3}$$

where we also assume the unobservable covariate has been defined so that $E[\omega_i^* \mid X_i]=1$ (this is a normalization since we can define $\omega = \pi^{-1}[\upsilon - \ln(E[e^{\upsilon} \mid X])]$, 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 non-negative fitted values:

$$E[Y_{1i} \mid D_i, Z_i, X_i, \omega_i] = e^{\alpha}E[Y_{0i} \mid D_i, Z_i, X_i, \omega_i] = e^{\alpha}E[Y_{0i} \mid X_i, \omega_i]. \tag{M4}$$

Combining (M3) and (M4) we can write

$$Y_i = exp(X_i'\beta + \alpha D_i + \pi\omega_i) + \varepsilon_i,$$

where $E[\varepsilon_i \mid D_i, Z_i, X_i, \omega_i] \equiv 0$. These assumptions imply

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

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 $e^{\alpha}-1$, which is approximately equal to α for small values of α.

Estimation based on (20) guarantees non-negative 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 non-linear 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 for this case, the proportional treatment effect can be obtained as follows:

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

In light of this simplification, it seems worth asking if the right hand side of (21) has an interpretation that is not tied to the constant-proportional-effects model. To provide 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} and D_{1i} as follows:

$$D_{0i} = 1[\gamma_0 > \eta_i],$$

$$D_{1i} = 1[\gamma_0 + \gamma_1 > \eta_i].$$

Using this notation, we have

$$E[Y_i|\, Z_i=1] - E[Y_i|\, Z_i=0] = E[Y_{1i}-Y_{0i}|\, D_{1i} > D_{0i}] \bullet \{E[D_i|\, Z_i=1] - E[D_i|\, Z_i=0]\}. \quad (22)$$

The term $E[Y_{1i}-Y_{0i}|\, D_{1i} > D_{0i}]$ is the LATE parameter mentioned in Lemma 1, in this case, for a model without covariates.

The same argument used to establish the original LATE result can also be used to show a similar result for the average of Y_{0i} (i.e., instead of the average $Y_{1i}-Y_{0i}$; see Abadie, 1998, for details). In particular,

$$E[(1-D_i)Y_i|\, Z_i=1] - E[(1-D_i)Y_i|\, Z_i=0] = -E[Y_{0i}|\, D_{1i} > D_{0i}] \bullet \{E[D_i|\, Z_i=1] - E[D_i|\, Z_i=0]\}. \quad (23)$$

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

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

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 M1-M4. Moreover, the exponential model used in M3 to incorporate covariates seems natural, and has a semi-parametric flavor similar to proportional hazard models for duration data.

4.2 Approximating causal models

Now, suppose that the additive, constant-effects assumptions (17a) and (17b) do not really hold and 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 paper, Abadie (1999) 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:

$$Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i \mid X_i, D_{1i}>D_{0i}. \tag{25}$$

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

$$E[Y_i \mid X_i, D_i=1, D_{1i}>D_{0i}] - E[Y_i \mid X_i, D_i=0, D_{1i}>D_{0i}] = E[Y_{1i}-Y_{0i} \mid X_i, D_{1i} > D_{0i}].$$

For this reason, Abadie (1999) calls $E[Y_i \mid X_i, D_i, D_{1i}>D_{0i}]$ the Complier Causal Response Function (CCRF).

Now, consider choosing parameters b and a to minimize

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

or, equivalently,

$$E[(Y_i - X_i'b-aD_i)^2 \mid D_{1i} > D_0].$$

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

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

where

$$\kappa_i = 1 - D_i(1-Z_i)/(1-E[Z_i \mid X_i]) - (1-D_i)Z_i/E[Z_i \mid X_i].$$

Since κ_i can be estimated, the MMSE linear approximation to the CCRF can also be estimated.

Note that while the above 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 chose parameters to minimize $E[\kappa_i(Y_i- \Phi[X_i'b-aD_i])^2]$. Similarly, for non-negative 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 zero, 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 impact of D_i is due to a pure participation effect 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 suggested that questions regarding the effect of treatment on the distribution of outcomes be addressed directly by comparing distributions. This is fine for the analysis of experimental data, but if 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:

$$1[Y_i \leq c] = X_i'\beta_c + \alpha_c D_i + \varepsilon_{ci}$$

Of course, the linear model is not literally correct for the conditional distribution except in special cases (i.e., a saturated regression parameterization).

Here too, the Abadie (1999) weighting scheme can be used to generate estimates that provide the 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 population minimand

$$E[\kappa_i(1[Y_i \leq c] - X_i'b_c - a_c D_i)^2]. \qquad (26)$$

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

$$P[Y_i \leq c | X_i, D_i=1, D_{1i} > D_{0i}] - P[Y_i \leq c | X_i, D_i=0, D_{1i} > D_{0i}] = P[Y_{1i} \leq c | X_i, D_{1i} > D_{0i}] - P[Y_{0i} \leq c | X_i, D_{1i} > D_{0i}].$$

Because the outcome here is binary, it also makes sense to consider a nonlinear model such as Probit or Logit to approximate $P[Y_i \leq c | X_i, D_i, D_{1i} > D_{0i}]$. Finally, note that it is equally straightforward to approximate the probability the outcome falls into an interval, instead of the cumulative distribution function.

An alternative to estimation based on (26) postulates a model for quantiles instead of distribution ordinates. Conventional quantile regression (QR) models 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 ordinates is application-specific. Second, since non-negative LDVs are often virtually continuously distributed away from the mass point, linear models are likely to be more accurate for conditional quantiles above the censoring point than for conditional probabilities. (For quantiles close to the censoring point, Powell's (1986b) censored quantile regression model may be more appropriate).

Abadie, Angrist, and Imbens (1998) developed a QR estimator for models with binary endogenous regressors. Their quantile treatment effects (QTE) procedure begins with a linear model for conditional quantiles <u>for compliers</u>:

$$Q_\theta[Y_i | X_i, D_i, D_{1i} > D_{0i}] = X_i' \beta_\theta + \alpha_\theta D_i.$$

The coefficient α_θ has a causal interpretation because

$$\alpha_\theta = Q_\theta[Y_{1i}|\, X_i,\, D_{1i}>D_{0i}]-Q_\theta[Y_{0i}|\, X_i,\, D_{1i}>D_{0i}].$$

In other words, α_θ is the difference in θ-quantiles for compliers.

The QTE parameters minimize 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 $\tilde{\kappa}_i$. Here I use the fact that

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

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, any negative estimates of $\tilde{\kappa}_i$ generated by the Probit first step are set to zero.

5. Application: The 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 percent file. About 53 percent 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 I, which reports descriptive statistics and repeats some of the OLS and 2SLS estimates from Angrist and Evans (1998). Roughly 38 percent of women in this sample had a third child, an event indicated by the variable *Morekids*. The OLS estimates 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 first-born, a dummy for male second-born, and black, hispanic, and other race indicators.

Table I also reports estimates of average effects computed using nonlinear models, still treating *Morekids* as exogenous. The average effects reported here are approximations to effects of treatment on the treated that use derivatives to simplify computations. A detailed description of the average effects calculations used for all the results appears in the appendix. The results of Probit estimation 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 that the *Tobit coefficient* is -11.7. This illustrates the importance of comparing apples to apples when using models like Tobit). Column 5 of Table I reports estimated average effects computed using a two-part model., where both parts of the model are linear. The 2PM estimate is virtually identical to the Tobit and OLS estimates.

Roughly 8/10 of one percent 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 6-7. The reduced forms show that women who had a multiple birth were 63 percentage points more likely to have 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 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 II, which report 2SLS estimates of a 2PM for hours worked and 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, possibly nonlinear, with some negative weights. Since use of negative weights is non-standard for statistical packages (e.g, Stata does not currently allow this), when computing Causal-IV estimates I used a MATLAB program written for this purpose by Alberto Abadie. The 2PM estimates in the table 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 since 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 are mostly similar to each other and to conventional 2SLS. Results from nonlinear models are reported as marginal effects that approximate average effects on the treated; For example, 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 percent larger than conventional 2SLS in absolute value. It is noteworthy, however, that the Mullahy model generates results that change markedly (falling by about 20 percent) when the covariates are dropped. Since the covariates are not highly correlated with the twins instrument, this lack of robustness to the choice of covariates seems undesirable. On the other hand, without covariates, the Mullahy estimates are close to those from exponential/Causal-IV.

The bivariate Probit estimate of the effect childbearing on employment status, reported in column 7, is -.12, roughly a third larger in absolute value than the conventional 2SLS estimate. Interestingly, in an

another application, Abadie (1999) 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 Probit first-stage (bivariate Probit, endogenous Tobit, and Mill's ratio), I randomly recoded 1 percent of D_i observations to zero. In principle, the resulting 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 II reports estimates of a structural Tobit model with endogenous 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 the censored regression, and then applies Tobit to the 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 a consequence of the Mills ratio procedure for controlling for selection bias in the estimation of treatment effects and not a consequence of the Tobit correction for non-negative outcomes. To see this, note that Mills ratio estimate of the effect of childbearing, 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. 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, parametric assumptions are weaker.

The last set of results show that childbearing is associated with marked changes in the distribution of hours worked. These results can be seen in the first columns of Table III, 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 II). 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 full-time. Once again, Probit average effects, reported in column 2, are almost indistinguishable from the OLS estimates. Estimates from an ordered Probit model, reported in column 3, differ from OLS somewhat more than simple Probit, 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 which 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 I. The 2SLS estimates of the impact of childbearing on full-time work are also considerably less than the corresponding OLS estimates.

As with results for mean outcome, non-structural Causal-IV models treating *Morekids* as endogenous generate estimates very close to 2SLS estimates of the effect 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 non-participation and by reducing the probability of full-time work. The quantile treatment effects estimator provides summary statistics for 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 leads to an estimated 9 hour decline in median hours worked, but the QTE

estimator shows that the causal effect of childbearing on median hours worked is only about 5 hours. Estimates at higher quantiles are similarly reduced when childbearing is treated as endogenous.

6. Summary and conclusions

Structural parameters must ultimately be converted into causal effects if they are to be useful for policy evaluation or determining whether a trend association is causal. The problem of estimating causal effects for LDVs 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 CEFs for LDVs in models with covariates. Without covariates, conventional 2SLS estimates capture both distributional effects and effects on means. Simple IV strategies adapted for nonlinear models can be used to estimate average effects in models with covariates, while IV strategies for probability models and quantile regression can be used to estimate effects on distributions.

These approaches are illustrated using twin births to estimate the labor-supply consequences of childbearing. Alternative non-structural approaches to IV estimation using twins instruments generate similar estimates, whether or not the model is nonlinear. Structural estimates tend to be somewhat larger 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, and cannot be computed without artificial modification of the data when using the twins instrument, I see no reason to prefer them. Finally, non-structural estimates of the effect of childbearing on the distribution of hours worked show that the impact of child-bearing is characterized by substantially increased non-participation and by an almost equally large shift away from full-time work. Estimates that treat childbearing as exogenous exaggerate the causal effect of childbearing on average hours worked and on changes in distribution. This finding is clear in results from both probability models and quantile models.

APPENDIX

Derivation of equation (12) in the text

Drop the *i* subscripts.

a. Model

$Y_0 = 1(Y_0^* > 0)\, Y_0^* \qquad = 1(Y_0^* > 0)Y_0^*$
$Y_1 = 1(Y_0^* + \alpha > 0)\,(Y_0^* + \alpha) \quad = 1(Y_1^* > 0)Y_1^*$

b. Causal effects

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:

$$E[Y_1 - Y_0 \mid Y_1 > 0, D=1] = E[Y_1 - Y_0 \mid Y_1 > 0]$$
$$E[Y_1 - Y_0 \mid Y_0 > 0, D=1] = E[Y_1 - Y_0 \mid Y_0 > 0],$$

so the effects on the right hand side are the parameters of interest.

c. Evaluation of expressions

(A) $E[Y_1 \mid Y_1 > 0] = E[Y_1^* \mid Y_1^* > 0] = E[Y_0^* \mid Y_1^* > 0] + \alpha$

(B) $E[Y_0 \mid Y_1 > 0] = E[Y_0^*\, 1(Y_0^* > 0) \mid Y_1^* > 0]$
$= E[Y_0^* \mid Y_1^* > 0,\, Y_0^* > 0]\, P(Y_0^* > 0 \mid Y_1^* > 0)$

$= (\alpha > 0):\ Y_0^* > 0 \Rightarrow Y_1^* > 0$, so (B) $= E[Y_0^* \mid Y_0^* > 0]\, P(Y_0^* > 0 \mid Y_1^* > 0)$
$(\alpha < 0):\ Y_1^* > 0 \Rightarrow Y_0^* > 0$, so (B) $= E[Y_0^* \mid Y_1^* > 0] \bullet 1$

So $\alpha < 0 \Rightarrow E[Y_1 - Y_0 \mid Y_1 > 0] = \alpha$

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

$= (\alpha > 0)$: $Y_0^* > 0 \Rightarrow Y_1^* > 0$, so $P(Y_1^* > 0 \mid Y_0^* > 0) = 1$
And $E[Y_0^* \mid Y_1^* > 0,\, Y_0^* > 0] = E[Y_0^* \mid Y_0^* > 0]$.

(D) $E[Y_0 \mid Y_0 > 0] = E[Y_0^* \mid Y_0^* > 0]$

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

d. Generalization to non-Tobit selection

Let W_0 and W_1 be potential outcomes that determine selection.

$$Y_0 = 1(W_0^* > 0)\, Y_0^* \qquad = 1(W_0^* > 0) Y_0^*$$
$$Y_1 = 1(W_0^* + \alpha_1 > 0)\,(Y_0^* + \alpha_2) \quad = 1(W_1^* > 0) Y_1^*$$

An analogous result holds: $\alpha_1 < 0 \Rightarrow E[Y_1 - Y_0 \mid W_1 > 0] = \alpha_2$

$\alpha_1 > 0 \Rightarrow E[Y_1 - Y_0 \mid W_0 > 0] = \alpha_2$.

Derivation of equation (21) in the text

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

$$\beta^* E[(1\text{-}D)Y \mid Z=1] + \beta^* \alpha^* E[DY \mid Z=1] - 1 = 0 \qquad \text{(A.1)}$$
$$\beta^* E[(1\text{-}D)Y \mid Z=0] + \beta^* \alpha^* E[DY \mid Z=0] - 1 = 0. \qquad \text{(A.2)}$$

Divide A.1 by A.2 to get rid of β^*, then solve for α^*. Subtract 1 and bring over a common denominator to get equation (21).

Average effects and standard errors for non-linear models

Average effects and standard errors were calculated with the aid of short-cuts and approximations that are likely to be useful in empirical practice.

Probit

The Probit average treatments effect were approximated using a derivative. Note that

$$\Phi[X_i'\beta+\alpha] - \Phi[X_i'\beta] \approx \varphi[X_i'\beta+\alpha D_i] \bullet \alpha ,$$

so the average effect on the treated can be approximated as

$$\{ (1/N_1) \textstyle\sum_i D_i \varphi[X_i'\beta+\alpha D_i] \} \bullet \alpha ,$$

where $N_1 = \sum_i D_i$. This approximation turns out to be accurate to three decimal places for the Probit estimates in Table I. Standard errors were calculated treating the scaling factor as non-random. 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 in Greene (1999):

$$E[Y_i|X_i, D_i=1] - E[Y_i|X_i, D_i=0] \approx \partial E[Y_i|X_i, D_i]/\partial D = \Phi[X_i'\beta+\alpha D_i] \bullet \alpha .$$

Average effects on the treated can therefore be approximated using

$$\{ (1/N_1) \textstyle\sum_i D_i \Phi[X_i'\beta+\alpha D_i] \} \bullet \alpha .$$

Standard errors were calculated treating the scaling factor as non-random.

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:

$$\alpha_1 E[Y_i| D_i=1, Y_i>0] + \alpha_2 P[D_i=1, Y_i>0].$$

Standard errors were calculated treating the scaling factors $E[Y| D_i=1,Y>0]$ and $P[D_i=1, Y>0]$ as non-random, 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

$$\omega_i^* exp[X_i'\beta+\alpha] - \omega_i^* exp[X_i'\beta] \approx \omega_i^* exp[X_i'\beta+\alpha D_i] \bullet \alpha$$

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:

$$\{(1/N_1) \sum_i D_i \, exp[X_i'\beta+\alpha D_i]\} \bullet \alpha \, .$$

Standard errors were calculated treating the scaling factor as non-random.

Bivariate Probit

Same as Probit, using parameters from the latent index equation for outcomes.

Endogenous Tobit

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 above 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 asymptotic formulas in Abadie (1999), and take account of the first-step estimation of $E[Z_i| X_i]$. As before, scaling factors were treated as non-random. Note that in contrast with the Mullahy estimator, there is no conceptual problem converting derivatives from the exponential Causal-IV model into average effects. An implicit assumption for all the Causal-IV models, however, is that it makes sense to convert conditional-on-X effects on compliers into effects on the treated.

Computation of Quantile Treatment Effects and standard errors

Quantile treatment effects 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. Non-negative 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 2500 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=2500, and N^* is the full sample size.

REFERENCES

Abadie, Alberto (1998), "Bootstrap Tests for the Effect of a Treatment on the Distributions of an Outcome Variable," Department of Economics, manuscript, December.

_____ (1999) "Semiparametric Estimation of Instrumental Variable Models for Causal Effects," MIT Department of Economics, manuscript, March.

Abadie, Alberto, Joshua Angrist, and Guido Imbens (1998), "Instrumental Variables Estimation of Quantile Treatment Effects," NBER Working Paper No. 229, March.

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, Joshua (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66, 249-88.

Angrist, Joshua and Guido Imbens (1995), "Average Causal Response With Variable Treatment Intensity," *Journal of the American Statistical Association* 90, 431-442.

Angrist, Joshua and Alan Krueger (1999), "Empirical Strategies in Labor Economics," Chapter 23 in O. Ashenfelter and D. Card, eds., *The Handbook of Labor Economics Volume IIIA*, Amsterdam: North-Holland.

Angrist, Joshua and William Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review* 88, 450-477.

Angrist, Joshua, Guido Imbens and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434):444-55.

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

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

Buchinsky, Moshe (1994), "Changes in the US Wage Structure 1963-87: Application of Quantile Regression," *Econometrica* 62, 405-458.

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

Cragg, John G. (1971), "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica* 39 [5] September, pp. 829-844.

Duan, Naihua, Willard G. Manning, Jr., Carl N. Morris, and Joseph P. Newhouse (1984), "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business and Economic Statistics* 2, 283-289.

Eichner, Matthew J. , et al (1997), "Health Expenditure Persistence and Feasibility of Medical Savings Accounts," in J. Poterba, ed., *Tax Policy and the Economy* 11, Cambridge, MA: NBER.

Evans, William, Matthew C. Farrelly, and Edward Montgomery (1999), "Do Workplace Smoking Bans Reduce Smoking?, *American Economic Review* 89 (forthcoming).

Gangadharan, Jaisri and Rosenbloom, Joshua L. "The Effects of Child-bearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5647,July 1996.

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

Greene, William (1999), "Marginal Effects in the Censored Regression Model," *Economics Letters* 64, 43-9.

Gronau, Reuben (1974), "Wage Comparisons-A Selectivity Bias," *Journal of Political Economy* 82, 1119-43.

Hay, Joel W. and Randall J. Olsen (1984), "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," *Journal of Business and Economic Statistics* 2 [3] July, 279-282.

Hay, Joel W., Robert Leu, and Paul Rohrer (1987), "Ordinary Least Squares and Sample-Selection Models of Health-Care Demand," *Journal of Business and Economic Statistics* 5 [4], pp. 499-506.

Heckman, James J. (1974), 'Shadow Prices, Market Wages and Labor Supply," *Econometrica* 42, 679-94.

_____ (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, James J. and Richard Robb, Jr. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data.* Econometric Society Monographs Series, No. 10. Cambridge: Cambridge University Press.

Imbens, Guido, and J.D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62 [2], pp. 467-475.

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

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger (1997) "Estimating Returns to Schooling When Schooling is Misreported" (unpublished).

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 G. Bassett (1978), "Regression Quantiles," *Econometrica* 46, 33-50.

Killingsworth, Mark (1983), *Labor Supply*, Cambridge: Cambridge University Press.

Keane, Michael P. and Kenneth Wolpin (1997) "Introduction to the *JBES* Special Issue on Structural Estimation in Applied Microeconomics," *Journal of Business and Economic Statistics* 15, 111-4.

Lee, Myoung-Jae (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, Siu Fai and Shihti Yu (1996), "On the Choice Between Sample Selection and Two-Part Models," *Journal of Econometrics* 72, pp. 197-229.

Lin, Tsai-Fen, and P. Schmidt, "A Test of the Tobit Specification Against and 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, pp. 3-18.

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

Manski, Charles 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, 709-33.

McDonald, John F., and R. A. Moffitt (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," Chapter 24 in O. Ashenfelter and D. Card, eds., *The Handbook of Labor Economics, Volume IIIA*, Amsterdam: North-Holland.

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

_____ (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics* 11, pp. 586-593.

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

_____ (1986) "Linear Instrumental Variable Estimation of Limited dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics* 32, pp. 127-141.

_____ (1987), "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics* 36, pp. 231-250.

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

Powell, James L. (1986b), "Censored Regression Quantiles," *Journal of Econometrics* 32, 143-55.

Rosenzweig, Mark R., and Kenneth Wolpin (1980), "Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models," *JPE* 88, 328-348.

Rubin, Donald 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).

Table I: Descriptive Statistics and Baseline Results

Dependent Variable	Mean (1)	*Morekids* exogenous: OLS Effect (2)	Probit (3)	Tobit (4)	2PM (5)	*Morekids* endogenous: Reduced-Forms (*Morekids*) (6)	(Dep. var) (7)	2SLS Effect (8)
Employment	0.528	-0.167	-.166	–	--	0.627	-0.055	-0.088
	(0.499)	(.002)	(.002)			(.003)	(.011)	(.017)
Hours Worked	16.7	-6.02	–	-6.01	-5.97	0.627	-2.23	-3.55
	(18.3)	(.074)		(.073)	(.073)	(.003)	(.371)	(.592)

Notes: The sample includes 254,654 observations and is the same as in Angrist & 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 first birth, boy first, boy second, and race indicators. Standard deviations are shown in parentheses in column 1. Standard errors are shown in parentheses in other columns.

Table II: Impact on Mean Outcomes (*Morekids* endogenous)

Dependent Variable	Linear Models: 2SLS (1)	2PM (2)	Causal-IV Linear (3)	Nonlinear Models: Mullahy (4)	Causal-IV Probit (5)	Causal-IV Expon. (6)	Structural Models: Bivar. Probit (7)	Endog. Tobit (8)	Mills Ratio (9)	2SLS Bench-mark (10)
A. With Covariates										
Employment	-0.088	--	-0.089	–	-0.088	–	-.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	-0.084	–	-0.084	–	-0.084	–	-0.086	–	–	–.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)

Notes: Sample and covariates are the same as Table 1. Results for nonlinear models are derivative-based approximations to effects on the treated. Causal-IV estimates are based on a procedure discussed in Abadie (1999). Standard errors are shown in parentheses.

Table III: Impact on the Distribution of Hours Worked

	Distribution Treatment Effects						Quantile Treatment Effects		
	Exogenous *Morekids*			Endogenous *Morekids*					
Range (mean)	OLS (LPM) (1)	Probit (row by row) (2)	Ordered Probit (3)	2SLS (4)	Causal-IV LPM (5)	Causal-IV Probit (6)	Quantile (value)	QR (7)	QTE (8)
0	0.167	0.166	0.147	0.088	.089	.088	0.5	-8.92	-5.24
(.472)	(.002)	(.002)	(.002)	(.017)	(.017)	(.016)	(8)	(.186)	(.686)
1-10	0.001	0.001	-0.002	-0.001	-.001	-.002	0.6	-12.7	-7.98
(.046)	(.001)	(.001)	(.00003)	(.007)	(.007)	(.006)	(20)	(.172)	(.860)
11-20	-0.015	-0.014	-0.011	0.002	.002	.002	0.7	-9.54	-6.19
(.093)	(.001)	(.001)	(.0001)	(.010)	(.010)	(.010)	(35)	(.184)	(1.07)
21-30	-0.024	-0.022	-0.014	-0.004	-.005	-.006	0.75	-6.45	-3.60
(.075)	(.001)	(.001)	(.0002)	(.009)	(.009)	(.010)	(40)	(.156)	(1.17)
31-40	-0.119	-0.110	-0.097	-0.072	-.072	-.071	0.8	-1.00	0.00
(.277)	(.002)	(.002)	(.001)	(.014)	(.014)	(.016)	(40)	(.286)	(1.09)
41+	-0.009	-0.008	-0.023	-0.009	-.009	-.006	.9	–	–
(.027)	(.001)	(.001)	(.0003)	(.005)	(.005)	(.007)	(40)		

Notes: The table reports probability-model and quantile treatment effect 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 II. Causal-IV estimates are based on a procedure discussed in Abadie (1999). Quantile Treatments are based on a procedure discussed in Abadie, Angrist, and Imbens (1998). Standard errors are shown in parentheses. The standard errors in columns 7 and 8 are bootstrapped.