Chapter 2
Issues in the Estimation of Causal Effects in Population Research, with an Application to the Effects of Teenage Childbearing

Robert A. Moffitt

2.1 Introduction

Population research has a distinguished history of empirical work on a wide variety of important topics related to population growth, the components of demographic trends, estimation of vital rates, life table construction, investigation into causes of historical population developments, and many others. However, one branch of population research that has seen increasing interest has been in the area of social demography, where the determinants of individual behavior regarding fertility, marriage, and related areas has been studied. It is in that branch that issues of causal inference have arisen, and with which this essay is concerned.

This development in population research coincides with a more general interest in causal inference in statistics and in other social sciences such as economics. In statistics, the development of the Rubin Causal Model (Rubin 1974) as a framework for studying causal questions has become a dominant paradigm even though, at the same time, there is considerable work by other statisticians using somewhat different frames. The development in statistics, while having many historical antecedents in the field, by and large occurred only in the 1970s and 1980s. Prior to that time, randomized experiments were regarded as the only method for true causal inference. However, randomized experiments are generally not possible in many fields, including population research, and hence methods for the analysis of causation using observational data are needed.

In economics, while causality has a much longer history dating to the development of the simultaneous equations model, which saw its fullest development in the Cowles Commission work in the 1950s, renewed interest in the issue has arisen since the 1980s and 1990s as more subtle issues have been addressed. Other social science disciplines such as sociology and political science are following the developments in statistics and in economics, with new developments adapted to their unique sets of questions and issues.

R.A. Moffitt (✉)
Department of Economics, John Hopkins University, 3400 N. Charles St., Baltimore, MD, USA
e-mail: moffitt@jhu.edu

H. Engelhardt et al. (eds.), *Causal Analysis in Population Studies*, The Springer Series on Demographic Methods and Population Analysis 23,
DOI 10.1007/978-1-4020-9967-0_2, © Springer Science+Business Media B.V. 2009

This essay will review the issues in causal modeling using primarily the framework adapted in economics, and will apply those modeling issues to the study of population questions. The economics framework is, when boiled down to essentials, observationally equivalent to the Rubin Causal model in statistics, although the interpretation and language used to describe the two are quite different. In addition, their practical empirical implementation is often quite different, with economists leaning toward modeling by the use of regression equations with explicitly represented error terms, an approach different from that in statistics.

While the causal modeling developments in economics have taken many directions, the vast majority of applications in the field use the method of instrumental variables (IV) to estimate causal effects. Therefore, this essay will also concentrate on that method, outlining both its rationale and advantages and the pitfalls and weaknesses associated with its use. Brief mention will be made of other methods such as panel data fixed effects methods and matching.

The running example in the essay is the question of whether teenage childbearing has a deleterious effect on female economic outcomes such as income and earnings. The increase in rates of teen childbearing in the U.S. has been a source of public concern not only because much of that childbearing is nonmarital but also because of the widespread perception that women who begin their childbearing at a very young age run the risk of harming their educational progress and their later economic and social success. There has been a great deal of research on this issue with, surprisingly, much less support for this conventional view as might be expected. But the literature has also generated much discussion of the method of causal effects and of the effects of using different instruments for estimation. Thus, this particular literature can be used to illustrate a number of the issues in causal modeling in population research in general.

The first section below lays out the general causal model in economics and discusses a number of the main issues. The method of instrumental variables is then outlined, followed by a categorization of the types of instruments most often used. Additional issues in the use of instrumental variables are then reviewed, followed by a set of conclusions. Some of the points made in the essay are (i) a tradeoff between internal validity and external validity is often faced by analysts using the method; (ii) multiple instruments or instruments with multiple values can be used to learn more about effects in heterogeneous populations than binary instruments; and (iii) use of theory is important to determine mechanisms by which treatments affect outcomes and how differing instruments interact with those mechanisms.[1]

2.2 The Basic Causal Model

The basic causal model in economics dates to the Cowles Commission work on simultaneous equations in economics and, later, its adaptation to individual actions represented in the switching regression model (Heckman 1978; Lee 1979).

[1] See Moffitt (2003, 2005) for earlier reviews.

Heckman and Robb (1985) and Björklund and Moffitt (1987) made the connection between that model and newer thinking in causal modeling as well as introducing the notion of heterogeneity to be discussed momentarily. Heckman et al. (2006) provide a recent overview of the model.

The prototype linear regression model used in this literature is

$$y_i = \alpha_i T_i + X_i \beta + \varepsilon_i \tag{2.1}$$

$$T_i^* = X_i \gamma + Z_i \delta + \upsilon_i \tag{2.2}$$

$$T_i = \begin{cases} 1 \text{ if } T_i^* \geq 0 \\ 0 \text{ if } T_i^* < 0 \end{cases} \tag{2.3}$$

where y_i is the value of the outcome for individual i, T_i is a dummy variable for whether an individual has received the "treatment," α_i is the effect of the treatment on y for individual i, X_i is a vector of exogenous covariates and β is its associated coefficient vector, Z_i is a vector of exogenous variables affecting the probability of receiving treatment but which do not affect y directly, and ε_i and υ_i are mean-zero error terms. This two-equation model, consisting of an outcome equation as a function of treatment and a second, selection equation (Equations (2.2) and (2.3) together) representing the determinants of treatment, is a special case of the general switching regression model. The selection equation is often not written down explicitly in the studies in the literature, and does not have to have the latent index structure shown in (2.2) and (2.3), but this is the most common interpretation. In addition, when estimating some of the objects of interest such as the marginal treatment effect or the local average treatment effect discussed below, an explicit representation of the selection equation is particularly helpful in interpretation.

The most important difference between this model and the classic linear simultaneous equations model in economics is that the effect of receiving treatment on the outcome (α_i) varies across individuals and hence treatment effects are "heterogeneous" in the population. In the older literature, this effect was assumed to be fixed. Allowing the effect to be heterogeneous has critical implications for interpretation and estimation. Most obviously, it requires a reconsideration of the object of estimation. The parameter α_i has a distribution in the population and one could imagine attempting to estimate different features of that distribution. One could attempt to estimate the mean of α_i, $E(\alpha_i)$, commonly called the average treatment effect, which is the average effect on y if all women in the population had a teen birth. Or one could attempt to estimate the average α_i for a subset of women observed, in a particular sample, to have had a teen birth. This object is $E(\alpha_i \mid T_i = 1)$ and is called the effect of the treatment on the treated. Two other possible objects of interest, the marginal treatment effect and the local average treatment effect, will be discussed below.

The assumptions on the covariance matrix of the error terms are that $E(\varepsilon_i \upsilon_i) \neq 0$ and $E(\alpha_i \upsilon_i) \neq 0$.[2] If y is individual earnings at some age like 25 and T is a dummy

[2] Conditioning on X and Z is left implicit.

variable for whether a woman had a teenage birth, $E(\varepsilon_i \upsilon_i) \neq 0$ if, for a group of women who have the same X, those women who had a teen birth would have had different future earnings than those women who did not, even if they had not had a teen birth. For example, those who had a teen birth may have come from disadvantaged backgrounds in unobserved ways. If, on the other hand, those who had a teen birth have a lower (i.e., less negative) impact of such a birth on future earnings than that of women who did not have a teen birth, then $E(\alpha_i \upsilon_i) \neq 0$. For example, women from disadvantaged backgrounds, who are more likely to have a teen birth, may be less affected by having a teen birth than women from less disadvantaged backgrounds because they have a lower payoff to human capital investments in the first place.

The implications for OLS of the two covariance assumptions are different. If $E(\varepsilon_i \upsilon_i) \neq 0$, OLS is biased for any object of estimation one might be interested in (i.e., any feature of the distribution of α_i).[3] OLS compares the y of women who had a teen birth to the y of women who did not, and this comparison is faulty because the y of women who did not have a teen birth does not equal the value of y for the women who did have a teen birth, if they had not had one. If $E(\alpha_i \upsilon_i) \neq 0$ but $E(\varepsilon_i \upsilon_i) = 0$, however, OLS returns a coefficient on T_i of $E(\alpha_i \mid T_i = 1)$, the effect of the treatment on the treated; but this is biased for the average treatment effect, if that is the object of interest. If both $E(\varepsilon_i \upsilon_i) \neq 0$ and $E(\alpha_i \upsilon_i) \neq 0$, again OLS produces nothing of interest.

Equation (2.1) is formulated as a linear regression function with additively separable X and with the vector of variables within X assumed to have effects through a linear index, $X\beta$. Thus, no nonlinearities in X or interactions between X and T are allowed. These restrictions can be relaxed by introducing such interactions into the model. Alternatively, a fuller nonparametric method could be used which allows T, and X to enter the model in arbitrary ways, although those methods typically lack power unless sample sizes are large. The method of matching is in this family, for matching is a method which nonparametrically estimates the effect of T on y, allowing X to affect y in an arbitrary fashion and allowing arbitrary interaction of T with X. However, the matching method is, explicitly or implicitly, an OLS method with an extended set of nonlinearities and interactions added, and hence requires that the error term in (2.1) be unrelated to υ_i (see Imbens (2004) for a review).[4] More formally, the assumption necessary for the method to produce unbiased estimates is the assumption of conditional independence, $E(\varepsilon_i \mid X_i) = 0$. Thus the method of matching is designed to address a different question than the standard selection model; the latter is designed to address the problem of unbiased estimation when selection is on unobservables, while the former is designed to address the problem

[3] OLS produces inconsistent and biased estimates because T is an "endogenous" independent variable, defined as one which is correlated with the error term in the equation even after controlling for X.

[4] That is, a fully nonparametric regression of y on X and T is equivalent to matching except that matching also imposes a "balancing" requirement to make the distributions of X in the $T = 0$ and $T = 1$ samples the same, whereas the nonparametric regression would not impose this.

of nonlinearities and interactions in the functional form of Equation (2.1) when selection is only determined by observables in the data (X).[5]

2.3 Instrumental Variables

The method of instrumental variables (IV) is designed to address the problem of selection on unobservables. It relies on the existence of the vector Z and, in its traditional two-stage least-squares form, simply involves first regressing T on X and Z and then regressing y on X and the predicted T from that first-stage equation. However, it can be equivalently represented in other forms which provide more intuition for what is being done and what is being estimated. To illustrate this point, let us simplify the model in (2.1)–(2.3) by omitting the control variables X and assume that Z is represented by a single binary (dummy) variable:

$$y_i = \beta_0 + \alpha_i T_i + \varepsilon_i \tag{2.4}$$

$$T_i^* = \gamma_0 + \delta Z_i + \upsilon_i \tag{2.5}$$

$$T_i = \begin{cases} 1 \text{ if } T_i^* \geq 0 \\ 0 \text{ if } T_i^* < 0 \end{cases} \tag{2.6}$$

For consistent estimation, the instrumental variable Z_i is assumed to have three properties: $E(\varepsilon_i \,|Z_i) = 0$, $E(\alpha_i \,|Z_i) = \bar{\alpha}$, and $E(T_i Z_i) \neq 0$.[6] The first two are "validity" restrictions that require that the instrument be mean-independent of the two unobservables in the y equation. The third is a "relevance" restriction which requires that the instrument be correlated with the probability of receiving treatment; the instrument must be "relevant."

When studying the effect of teenage fertility on later earnings, for example, the availability of contraceptives in a woman's geographic area might satisfy these conditions. For example, contraceptives might be more available in one area than another because of different governmental decisions to provide them. Those governmental decisions may be unrelated to the unobserved earnings levels of the women in each area (ε_i) or to the effects of teen childbearing on earnings of those women (α_i). The assumption would fail, however, if governments provided more contraceptives when the women in the area are more economically disadvantaged. Availability of contraceptives is likely to be relevant since it presumably affects teen childbearing.

A true experiment would guarantee that the conditions were met. Suppose that Z is a dummy for whether a women is randomly assigned to an experimental group or

[5] In principle, there is no reason that attention to nonlinearities and interactions in Equation (2.1) cannot be addressed at the same time as addressing selection on unobservables.

[6] These assumptions are stronger than what is needed and can be relaxed, but make the exposition particularly simple.

a control group. The experimental group is offered additional contraceptives while the experimental group is not (this is called an "offer" experiment; not all women in the experimental group use the contraceptives). This experiment should affect the treatment variable T, i.e., whether a birth occurs, and hence Z should be relevant. Because of the randomization, Z should be unrelated to the earnings functions of the women in the experiment in the absence of the treatment, i.e., the women in the experimental and control groups should be identical in all observed ways (ε_i and α_i). Thus Z should also be valid. The search for a valid and relevant Z in observational data is essentially a search for a "natural experiment" where Z is effectively randomly assigned to different groups within the population.

An important question is whether there is any way to test whether a particular variable is a suitable instrument and meets the criteria of validity and relevance. Relevance can be ascertained to some degree by determining how significant and strong a determinant of T a particular variable is, either by examining the t-statistic or F-statistic for the coefficient on Z when estimating T as a function of Z. Validity, however, cannot be tested if there is only one Z being examined (the case of a "just-identified" model). To do so requires that a sample estimate of the covariance between Z and ε (in a simple linear model) be obtained, and it is not possible to obtain a consistent estimate of ε without the assumption that Z is valid in the first place. Determining whether a variable is a valid instrument therefore requires appeals to theory and priori arguments for why Z is likely to be randomly assigned, usually based on arguments for why the determinants of Z are likely to be uncorrelated with the unobservable determinants of y, as well as empirical investigations into the observable determinants of Z which, while not constituting proof of its lack of correlation with unobservables, nevertheless can give clues to the process of its determination.

When Z is a dummy variable, the two-stage least-squares estimation of the model yields a coefficient on T in the y equation equal to:

$$\widehat{\alpha}_{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{T}_1 - \bar{T}_0} \tag{2.7}$$

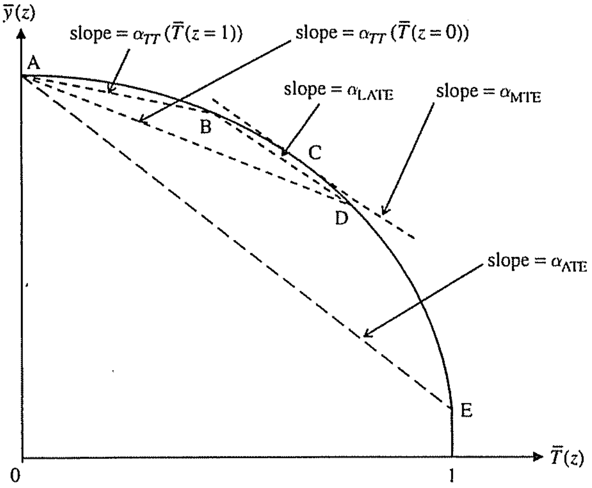
where $\bar{y}_1$ is the mean of y_i over all the observations for which $Z_i = 1$, $\bar{y}_0$is the mean of y_i over all the observations for which $Z_i = 0$, $\bar{T}_1$is the mean of T_i over all the $Z_i = 1$ observations, and $\bar{T}_0$ is the mean of T_i over all the $Z_i = 0$ observations. This formulation of the IV coefficient, emphasized by Imbens and Angrist (1994) and termed the local average treatment effect (LATE) for reasons to be explained below, is instructive in understanding how IV effects are estimated. In the case of a binary Z, the numerator of (2.7) represents the difference in the value of y for those in the "experimental group" ($Z = 1$) and those in the "control group" ($Z = 0$). In each of these groups, the mean of y is a weighted average of those with $T = 1$ and $T = 0$ in the group. The denominator represents the difference in the fraction of each group which "participates" by "taking up" the offer of treatment.

While Equation (2.7) involves only overall means in the two groups, it can be interpreted as representing the change in y for the subset of observations who changed their value of T as a result of the differing Z. For example, if contraceptives are more available in one area ($Z = 1$) and less available in another ($Z = 0$), the denominator represents the difference in the fraction of women who have a teen birth because of greater contraceptive availability. If the instrument is relevant, this change will be nonzero, that is, contraceptive availability affects teen birth probabilities. The numerator represents the difference in mean earnings in the two areas and is assumed to be solely a result of the change in teen births; this follows from the assumption of validity, which implies that earnings in the two areas are identical aside from Z and hence any difference in earnings can be ascribed to the effects of differing Z. The change in earnings has to be "inflated" by the change in the teen birth fractions between the areas. So, for example, if the difference in average earnings between the areas is $1,000 per year and the change in the teen birth fraction as a result of increased contraceptive availability was -0.10, then the change in y for the 10 percent of women who changed their birth behavior must have been $10,000. That is because 90 percent of women did not change their value of T at all, and hence must have had no change in their y.

This formulation of the IV method when the instrument is binary underscores the limited nature of what has been learned. The effect being estimated is the average effect of treatment, such as having a teen birth, for the 10 percent of women who switched from not having a teen birth to having a teen birth (called "switchers" or, in the language of Angrist et al. 1996, "compliers"). The effect of having a teen birth on outcomes cannot be assumed to be the same for any other group. For example, if the difference in contraceptive availability in the two areas in the data changes the teen birth rate from 0.30 to 0.20, its effect on earnings cannot be assumed to equal what would happen if the birth changed from 0.20 to 0.10, or for any other rate. If α_i is heterogeneous, as is assumed here, then the effects of teen birth on earnings will differ for different groups. If, for example, those women who have teen births when there are few contraceptives available include many for whom the negative consequences on earnings are particularly large, then expanding contraceptive availability and lowering the birth rate may have a gradually smaller and smaller effect on earnings, as those who continue to have birth rates even when contraceptive availability is high are those who have the smallest negative consequences for earnings.

This case is illustrated in Fig. 2.1. The mean of earnings in an area is plotted against the teen birth rate in the area as line ABCDE. As the birth rate rises, mean earnings fall as a larger fraction of the population has a reduction in earnings as a result of a teen birth. However, the figure is drawn such that the slope of the curve increases (becomes more negative) as the fraction with a teen birth rises, as would be the case if those who have the largest negative earnings reductions are the "last" ones to have a teen birth. The LATE estimate of the effect of teen births on earnings when the binary instrument lowers the teen birth rate from $\bar{T}(Z = 0)$ to $\bar{T}(Z = 1)$ is the slope of the dotted line connecting points B and D. It is termed the local average treatment effect because it applies only to a "local" area of the curve (that between B and D) and is only an average of those who change teen birth behavior in that region. Obviously, a LATE estimate will differ depending on where the two points

Fig. 2.1 Treatment effects



induced by the instrumental variable are on the curve and, in this sense, there is no longer a single effect of treatment on outcomes; one cannot speak of "the" effect of teen births on earnings, for example, for "the" effect depends on the population affected. The LATE estimate will not differ across ranges of fractions treated only if the curve in Fig. 2.1 is a straight line, with constant slope.

The other two possible objects of interest mentioned above, the average treatment effect (ATE) and the effect of the treatment on the treated (TT), are also shown in the figure. The ATE is the slope of the line connecting points A and E, representing the change in earnings that would occur if the entire population went from no teen births to all teen births. Two values of the TT are shown by the slopes of the dotted lines AB and AD, which show the difference in earnings that would arise if the fraction of women having a teen birth in the population were $\bar{T}(Z = 1)$ or $\bar{T}(Z = 0)$, respectively, rather than no teen births. The IV estimate obtained from a binary instrument Z will not equal a TT unless one is lucky enough to have an instrument for which $\bar{T} = 0$ (e.g., an area where contraceptives are so available that the teen birth rate is almost zero) and will not estimate the ATE unless one is even luckier to have an instrument for which the teen birth rate is zero for some values of the instrument and is equal to one (everyone has a teen birth) for other values of the instrument. In most applications, this is extremely rare.

Learning only the effect of the treatment in a local area may not be limiting if the other areas of the curve are not particularly relevant. No populations or subpopulations defined by demographic groups have teen birth rates close to 100 percent, so the lack of knowledge in that region of the population may not be discouraging. If the instrument in question moves the teen birth rate over a region which is very similar to that of most other populations and periods one is interested in, it may not be very disadvantageous.

In some cases, the researcher may be willing to extrapolate beyond the data available and draw conclusions about the effect of treatment on other populations.

Such extrapolation is a special case of the problem of "external validity" originally discussed in the context of classical experiments, where the question is whether the results of an experiment can be generalized beyond the specific type of program examined and beyond the specific type of population enrolled in the experiment. Extrapolation is always possible, but it is necessary to be clear that to do so requires additional assumptions (e.g., on the shape of the curve outside the range of the data) than were necessary to obtain the initial estimates, and a clear separation between the two needs to be made.

Learning the effect of teen births over more points on the curve requires more instruments or more variation in a single instrument. Sometimes this can arise across studies using different instruments, in which case one could imagine piecing together the curve from different investigations. Alternatively, if multiple instruments or instruments with more than a two values are available, the effects can be estimated over greater portions of the curve. Carneiro et al. (2006) and Moffitt (forthcoming) have considered estimation in these circumstances, both noting that Heckman and Vytlacil (2005) showed that the curve in Fig. 2.1 can be represented by the regression equation

$$y_i = X_i\beta + g\left[T\left(Z_i\right)\right] + \varepsilon_i \tag{2.8}$$

where a vector of other variables (X_i) has been reintroduced. The function g is intended to represent the curve ABCDE in Fig. 2.1, and can be specified as a polynomial, piecewise-linear function, or estimated completely nonparametrically.[7] To estimate it requires a first-stage estimate of $T(Z_i)$, which is the same as the first-stage estimation in traditional two-stage least-squares estimation of these models. The predicted probability of $T = 1$, as a function of Z, is then inserted into (2.8) and estimation can proceed.

Heckman et al. and Moffitt, following on the terminology of Björklund and Moffitt (1987) and Heckman and Vytlacil (2005) term the slope of the "g" function the "marginal treatment effect" (MTE). It is simply the slope of the curve in Fig. 2.1, and will vary over the range of $\bar{T}$. The slope of the dotted line at point C in Fig. 2.1 shows the MTE at that point, and Fig. 2.2 shows the MTE over the entire range. The MTE represents the effect of the treatment on the outcome for the marginal person "brought into" treatment by a small increase in the fraction treated. Estimates of Equation (2.8) may reveal, on the contrary, that y is linearly related to $T(Z)$, in which case the curve in Fig. 2.1 is a straight line and it can be concluded that there is no heterogeneity of response in the population.

As before, how much can be learned from such an exercise depends on the range of $\bar{T}$ induced by the range of instrumental variables in the data. If the entire range of $\bar{T}$ from 0 to 1 is not induced, only a portion of the curve can be estimated. Nevertheless, the general lesson is clearly that instruments which induce a wider

[7] Carneiro et al. proposed applying the partial-linear regression model to obtain nonparametric estimates, while Moffitt proposed applying series methods.

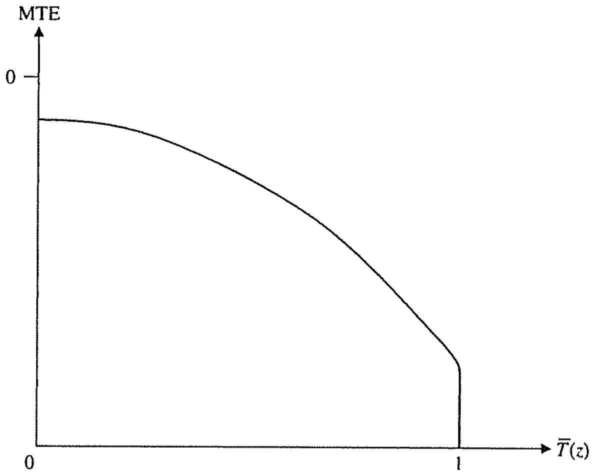


Fig. 2.2 Marginal treatment effect

range of fractions treated are more desirable than those which induce a smaller range.

If multiple instruments or instruments with more than two values are available, application of IV (e.g., in its traditional two-stage least squares form) to Equation (2.1) without allowing for nonlinearity of $T(Z)$ will produce a single coefficient which is a weighted average of the different MTE's over the range of the data (Angrist and Imbens 1995; Angrist and Krueger 1999; Heckman and Vytlacil 1999). This weighted average can be more difficult to interpret than either the LATE for a binary instrument or the varying MTE estimated in the Heckman-Moffitt approach because it is unclear what range of population it applies to. Nevertheless, it roughly characterizes the average effect in a loose sense.

2.4 Types of Instrumental Variables

The range of types of instrumental variables used in the literature in economics, demography, and other fields is very wide, and hence any attempt to group them into different types must necessarily be only approximate. But with this caveat in mind, the large majority of instruments can be classified into one of four types: cross-sectional ecological variables including area fixed effects instruments, population-segment fixed effects instruments, sibling and related instruments, and a residual category of "natural experiment" variables.

Cross-sectional ecological variables are variables which affect the environment in which individuals make choices and are most often measured at the aggregate level, most commonly for a geographic area. The variable for availability of contraceptives discussed in the teen birth case is an example of this type of instrument (Klepinger et al. 1999). Similar instruments used in this literature are state restrictive abortion laws and state family planning services (Klepinger et al. 1999) and the availability of gynecologists in the individuals' area (Ribar 1994). More generally, variables measuring differences in laws, labor markets, social structure, prices and availability of services (e.g., child care) in an area are used. The argument for the validity of such instruments is that they are "external" to the individual's own behavior, and therefore can be argued to be unrelated to the individual's individual determination of his or her outcome, y. Unlike individual-level characteristics such as family background or related measures, for example, which are quite likely to be direct determinants of y and hence not excludable, the higher-level instruments could be argued to not directly appear in the y equation.[8]

The most common objection to ecological variables is the well-known problem of unobserved ecological correlations. One common type of correlation arises when individuals who live in different areas are different in unobserved ways, as might arise through residential sorting. Another is that, even if residential sorting does not take place, unobserved area-level factors may be present which cause differences in individual outcomes across areas, even holding individual characteristics fixed. However, both of these problems cause difficulties only if the unobserved differences in question are correlated with the area-level instrument being employed. To consider this requires investigating the determinants of the instrument and why its value varies across areas. For example, if the availability of contraceptives is partly affected by the level of teen births in an area, implying that any unobservable affecting teenage fertility is correlated with contraceptive availability, the instrument will be invalid. In many cases, the political process (e.g., in the case of laws) has to be considered and, in some studies, a rather detailed study of the reasons for passage of a particular piece of legislation is provided in order to prove that the reasons for passage were purely "political" and not related to the value of y in the area.

In many cases, objections to the validity of ecological instruments on the basis that those instruments are correlated with area-level unobservables can be addressed if the instruments change in value over time differently in different areas, and if data are available on samples of individuals over time as well. In this case, an area fixed-effects model can be estimated, most simply by estimating the model in first-difference form.[9] In this case, the main equation is formulated as one in which Δy is assumed to be a function of ΔT, and the instrumental variable used is ΔZ.[10] Estimation in this form will eliminate any area-level unobservables that are fixed over time, which will be differenced out. For example, examining how the change in earnings over time across areas is related to the change in contraceptive availability,

[8] The multi-level model in social statistics bears a relationship to ecological instruments but is aimed at a different problem, for the multi-level model is aimed at obtaining correct and efficient standard errors rather than addressing the problem of endogeneity of a treatment variable.

[9] As in the standard fixed effects model, however, the "within" estimator is more efficient than the first-difference estimator.

[10] If panel data are available, this model can be estimated directly. If only repeated cross-sections are available, the model has to be formulated slightly differently, with the waves of the data pooled and an interaction term between time dummies and T and Z entered.

working through changes in teen birth rates, may be a more reliable method. Once again, however, one must carefully consider why contraceptive availability changed in different ways across areas, to insure that it did not change in response to trends in the teen birth rate in the area in question.

It is worth noting that estimation of the panel data fixed effects model without adjustment for endogeneity with instrumental variables is no longer favored within economics, whereas it was initially thought by some researchers to be an acceptable solution to the problem. Simply examining whether a change in an independent variable (T) for an individual is correlated with a change in outcome (y), which will eliminate individual-level fixed effects, is now thought to problematic because individuals change their actions (T) over time for reasons that are usually related to changes in their situation that are also affecting y. It is still necessary to find some determinant of the change in T that can be more plausibly argued to be unrelated to the forces at the individual level that are driving changes in y.

A second type of instrument can be termed, for lack of a better term, the population-segment fixed effects variable. In this case, changes in outcomes y for two groups which experience different changes in T and Z are compared, and the difference in those changes are assumed to be a result of the changes in T and Z which also occur over time. The difference with the area fixed effects model is that the groups are not defined by geographic location but rather by demographic or economic group. For example, suppose that contraceptives are more or less freely available to all higher-income and more-educated groups at all times, but become more available to lower-income families over time (in the nation as a whole; no geographic variation is assumed to occur). In that case, the variation in the change in Z over time arises from differences by income or education group. The validity of the instrument depends on the accuracy of the assumption that both groups would have had the same change in their earnings (controlling for changes in observables, X) in the absence of a change in contraceptive availability for the lower-income, less-educated group. Put differently, the assumption for validity is that the earnings of both groups are trending at the same rate in the absence of the change in Z. This may not be the case if there are other time-trending unobservable factors affecting the two groups differently.

There have been no applications of this method in the teenage birth case, but it has been used frequently in studies examining nationwide changes in governmental policy such as changes in the welfare system (Moffitt and Ver Ploeg 2001, Chapter 4). There, changes in earnings, fertility, or other outcomes over time for groups primarily affected by welfare reform (e.g., less-educated single mothers) are compared to those changes for groups presumably unaffected or less affected (more educated single mothers, single childless women, married mothers, men). The assumption that the different demographic groups, or population segments, trend at the same rate in terms of earnings, fertility, and other outcomes would appear to be a very strong one both because so many other social, economic, and political forces are typically changing over time that may affect the groups differently but also because, in some cases, the policy in question may affect the characteristics that define the groups (e.g., marriage or fertility).

A third type of instrumental variable used frequently in recent years is based on sibling or twin differences. Assuming data are available on a sample of individuals, some of whom are twins or siblings, Equation (2.1) can estimated on the pooled sample of all individuals. The instrument in this case is $(T_{if} - \bar{T}_f)$, where T_{if} is the treatment value for individual i in family f and $\bar{T}_f$ is the average of T_{if} over all individuals in family f. Thus the instrument is the deviation of each individual's T from the family-specific mean. The assumption needed for validity in this case is that that deviation is independent of the deviation of each individual's ε from its family-specific mean. In the teen childbearing case, for example, the necessary assumption is that the fact that one sibling has a teen birth and another does not is unrelated to their future earnings or any determinant of future earnings such as ability, motivation, or other factors related to economic success. What the mean eliminates as a source of problem are the differences that arise from different levels of disadvantage across families, which is almost surely related both to teen childbearing and to later earnings. Nevertheless, the assumption remains a very strong one, for there are well-known differences in sibling development and in parental treatment of siblings that could cause the necessary assumption to fail (for a discussion of these issues by economists, see Bound and Solon (1999)).

Even though the model can be estimated by IV, most often it is estimated instead in reduced form. Assuming that Equations (2.2) and (2.3) are "substituted" in for T_i in Equation (2.1) (ignoring the nonlinearity involved in the substitution), a reduced form equation is obtained specifying y_i as a function of X_i and Z_i. Using the within-family deviation on T for Z, OLS estimation of such an equation is equivalent to estimating the model in within-family differences (that is, regressing the within-family deviation of y on the similarly-defined X and Z). Geronimus and Korenman (1992) were the first to apply this method in the teen birth literature, and found it to have a large effects on the results relative to OLS. Hoffman et al. (1993) and Ribar (1999) have provided further discussion of the method and the results.

A fourth category of instrument, really a residual category comprising several different types of approaches, is the use of "natural experiments" as instruments. The term refers to occurrences of "random" events that arise in "nature" (that is, not in a controlled laboratory setting) which can be arguably unrelated to unobservables in many individual outcomes (Angrist and Krueger 2001). In fact, defining the term this broadly would include all the other instruments already discussed, so the term in this case is more narrowly defined. One type included here is really a subset of the area fixed effects model and is applied whenever there is a law or policy change that applies to a narrow group of the population in very similar circumstances. For example, a law which affects only children in a particular age range and in a particular income in one state but not another states could be arguably unrelated to various later outcomes for the children (see Currie and Gruber (1996) for a related example). This differs from the general area fixed-effects model only by virtue of using a much narrower demographic and geographic segment of the population. Another category is what Rosenzweig and Wolpin (2000) have called "natural natural experiments" which arise when a possibly random demographic event occurs such as the birth of twins, the month of the year in which a child is born, or whether a miscarriage occurs for a woman who is pregnant. The birth of twins has been argued to cause a random increase in the number of children and hence may be arguably used to estimate the effects of childbearing on a wide range of outcomes (e.g. Angrist and Evans 1998). The month in which a child is born has been argued to affect when a child can enter school and when a child is legally eligible to drop out of school, and hence is arguably a random determinant of years of education (Angrist and Krueger 1991). Miscarriage can be argued to also randomly affect fertility, or at least its timing, and has been used as an instrument for the likelihood of a teen birth (Hotz et al. 1997a, b). The age at menarche is another possibly random variable which should affect the age at which a first birth can occur and hence the probability of a teen birth (Ribar 1994; Chevalier and Viitanen 2003). Yet another type of instrument which may be termed a natural experiment are instruments based on so-called regression discontinuity designs (Cameron and Trivedi 2005; Imbens and Lemieux, forthcoming). This approach makes use of cases where there is an important variable which affects outcomes, and therefore is not excludable, but there also exists a policy or other event which generates a discontinuous change in that variable at a discrete point in the range. For example, if a family planning organization provided free contraceptives and other services strictly only to those with incomes below the poverty line, then a comparison of the outcomes of families just above and just below the poverty line should come close to measuring the effects of free contraceptives, other things being equal, because the two groups of families are "almost" identical in terms of income.

The validity of natural experiment instruments must be considered on a case-by-case basis, for there are very few generalizations possible. In the case of differing laws across states, the same issues have to be considered as in the general area fixed-effects model. For the natural natural experiments, threats to validity are often based on the suspicion that the demographic event in question has direct effects on the outcomes of interest. For example, a miscarriage may affect the mother's attitudes toward future fertility or her educational and economic outcomes directly, and not simply through the consequent postponement of fertility. There is also the possibility that miscarriage is related to underlying physiological or health factors that might be related to later economic success. Likewise, age at menarche might be related to unobserved health factors. Some demographers have argued that month of birth is correlated with other variables affecting the timing of fertility and the probability of conception, and therefore may have some indirect correlation with later educational outcomes, for example.

A rather different concern with natural experiments is that they are rather limited in their external validity. The narrow-population law differences necessarily apply only to a small segment of the population, for example, whose effects may not generalize to other parts of the population or to other laws. The regression discontinuity design necessarily can estimate impacts only for those around the point of discontinuity, e.g., only those with incomes just around the poverty line. Again, those effects may not generalize to other groups in the population. These issues of external validity are in addition to those discussed in the prior section related to varying MTE over the range of fractions population treated; the issue here is simpler and more traditional because the groups whose effects are estimated are characterized by standard socioeconomic observables (age of the child, family income, etc).

The issue of external validity raised in the natural experiment literature also serves to illustrate a seeming tradeoff between internal and external validity. Internal validity, defined in this case to be the validity of an instrument in an IV context, is attempted in the natural experiment literature by focusing on narrowly defined groups for whom it is plausible not only that observables are equivalent but also unobservables. But maximization of internal validity comes at the cost of sacrificing external validity, i.e., generalizability. It is possible that a "partially valid" instrument applied to a larger population could generate estimates which are biased but have acceptable mean-squared error and yet are more generalizable. This issue is difficult, if not impossible, to address because it is not clear how this tradeoff can be formalized. As a result, much of the literature, particularly that on natural experiments, has moved toward maximizing internal validity to at least learn something definite even if on a narrow population.

2.5 Additional Issues

The discussion of types of instruments provided in the last section is sufficient in and of itself to illustrate many of the issues with instrumental variable methods for estimating causal effects. Some of these issues bear further discussion and emphasis, and some additional issues should be introduced.

2.5.1 Heterogeneity

The issues discussed earlier in the paper surrounding the existence of heterogeneous effects and varying MTE, and the importance of identifying the specific population and point on the response curve from which estimates are being generated, has not penetrated most of the applied IV literature at this writing. In a large majority of the cases, binary instruments are used, which only permit a single LATE estimate to be obtained. Only rarely are comparisons made across studies in an attempt to piece together a fuller picture of the entire response curve (Card (2001) is an example of such an attempt). In addition, even in cases where multiple-valued instruments or multiple instruments are used, typically a single IV coefficient is estimated when more could be learned by estimating some portion of the MTE response curve. The literature is still young in this regard.

2.5.2 Differences in Estimates Across Instruments

One of the more difficult issues in the literature is that different instruments appear to generate differences in estimates for reasons not apparent. One possible reason is that just noted, that different instruments are moving the fraction of the population that is treated across different ranges of the population response curve. However, this often does not appear to explain the main differences in effects across instruments.

For the teen childbearing case, for example, Reinhold (2007) conducted an investigation into the reasons for differences in the effects of teen childbearing on high school completion using miscarriage as an instrument, which implies that childbearing has no effect if not a positive one, versus age at menarche, which implies a negative effect (albeit one smaller than OLS). Reinhold found that the differences could not be explained by the heterogeneous response curve, for the instruments yielded different estimates even when estimated at the same point on the curve (although the estimates also had large standard errors). Reinhold's investigation revealed that miscarriage occurs disproportionately among very young women, possibly because of immature physical development at that age, and that a comparison of women who miscarry with those who do not involves a comparison of women with different ages at the time of pregnancy. If those who become pregnant at young ages have more negative effects of teen childbearing on completed education than those who become pregnant at older ages, this lack of proper control for age could lead to small or zero effects of teen childbearing on education when using miscarriage as an instrument.

More generally, a possible cause of differences in IV estimates for different instruments is that they affect different types of individuals. Miscarriage affects women who have expressed a desire to have a child when young (assuming the child is desired) whereas increased contraceptive availability is likely to affect women who desire to have a child later and want to avoid having one early. These two types of women may not be affected in the same way by a postponement of childbearing, although it is not clear on a priori grounds which group would have the greater impact on later earnings. Similar reasoning could apply to other instruments.

A somewhat related reason for different estimates across instruments is that different instruments may work through different mechanisms, or channels. Indeed, the effect of teen childbearing on later earnings can have many different pathways. It could affect educational outcomes, which in turn affect earnings; the presence of young children could affect the ability to work after leaving school, even if education is unaffected; it could affect the types of jobs that a woman can and is willing to take, for similar reasons; or it could affect the probability of marrying, and marital status is known to have a significant effect on labor force participation and earnings. The increased control of fertility made possible by more availability of contraceptives, for example, may allow a woman to proceed with marriage to a suitable partner, knowing that fertility can still be controlled. Miscarriage may not have the same effects on marriage probabilities if fertility is relatively uncontrolled, and marriage may be delayed. To some extent, these differences in instrument effects can be examined by studying different mechanisms, of course, but this is only occasionally done.[11]

[11] In all of these cases, the "compliers" – that is, the women whose decisions are changed by the variation in the instrument – are different. See Angrist and Imbens (1995) for a discussion of IV with multiple instruments.

That the particular mechanism through which a treatment has an effect may matter for IV estimation is an illustration of the more general point of Rosenzweig and Wolpin (2000) that a careful specification of theory is necessary prior to applying instrumental variables. They provide several illustrations where the interpretation of IV estimates differ dramatically depending on what other variables are controlled for in the equation and by what channels treatment has an effect.

2.5.3 Relevance of Instruments to Policies of Interest

Yet another issue of some importance in the IV literature is whether the instruments used are relevant to the policies or programs that might be the ultimate goal of the analysis. This problem appears most often with experiments of the "natural natural" type, where twins, month of birth, miscarriage, age at menarche, and related variables are used. These variables do not directly relate to any public policy and hence may be difficult to use to learn what the effects of policies might be. This is particularly true when there are multiple mechanisms by which the treatment can affect outcomes, for particular government policies may work in different ways. If teen childbearing were addressed by providing extra subsidies to stay in school, for example, that could have a different effect on later earnings than the effect of a miscarriage.

Studies which use policy-based instruments are preferable from this point of view because they provide direct information on the effects of at least one concrete policy. Even here, however, it is unclear how to use the results of a policy instrument study to forecast the effects of some other government policy which works in different ways. Both of these examples suggest that some attention should be paid to the specification of the selection equation and to using theory to account for the different mechanisms by which variables affect the treatment. For example, the monetary cost of staying in school is, according to economic theory, one possible determinant of voluntary teen childbearing. A study which represented that cost explicitly in the teen birth probability equation and which estimated its effects would allow a "mapping" of the effect of the instrument into the effect of schooling costs, and thereby permit an estimate of the effect of such alternative policies.

2.5.4 Reduced Form Versus Structural Form

Some of the studies in the IV literature estimate reduced forms rather than structural forms; that is, they estimate models for y as a function of X and Z directly. In the case of a binary instrument, this effect is simply the numerator of Equation 2.7 and therefore simply equals the IV coefficient multiplied by the change in the fraction of the population treated, so there is no real difference between them. However, estimation of the reduced form alone is generally conducted only when the instrument in question is of direct policy interest, for the results of reduced form estimation will yield an estimate of the effect of the policy even if it is does not work through the particular T specified in the model. On the contrary, simply estimating the effects of, say, having twins on later female earnings is not in and of itself very interesting unless it is interpreted as working through effects on family size. Nevertheless, even in the former case, most analysts believe that estimation of the structural form is of the greater theoretical interest because learning the mechanism by which policies have their effects, and that knowledge of this mechanism is necessary to design new policies which work through the same mechanism.[12]

2.5.5 Weak Instruments

The criterion of relevance for an instrument is an important one, for in many cases an instrument, while having a large asymptotic t-statistic in the estimation of the effect of the instrument on T, may nevertheless have small explanatory power for T. In that case, the instrument is said to be "weak" and the IV estimate of the effect of T on y can be shown to be biased toward OLS, and to have much larger confidence intervals than produced by the usual formulas (Cameron and Trivedi 2005). Rules of thumb have been developed for detecting weak instruments based on the F-statistic for a single instrument in the T equation, e.g., that it should be at least 10 (Stock et al. 2002; Stock and Yogo 2005) as well as more formal methods, and there are also formal methods for calculating more accurate confidence intervals for effect estimates when instruments are weak. The implication for practice is that a somewhat higher standard for instruments must be applied, for they not only must be valid in the standard asymptotic sense, but they must be sufficiently "strong." In many applications, instruments which are arguably exogenous on theoretical grounds or which appear to have a significant coefficient in the T equation nevertheless are only weakly related to the fraction of the population treated, in which case usually a search for a stronger instrument is required.

2.5.6 No Instruments Available

The thinking about instrumental variables described in this essay has led to a higher standard for the choice of instruments than existed in earlier years of research, when the attitude toward instruments was more casual. This has made the search for a suitable instrument more difficult and in some cases no credible instrument exists either conceptually or in the available data sources. Particularly if attention is restricted to pure natural experiments of the type described above, the relative infrequency of

[12] Heckman and Vytlacil (2001) have argued that the reduced form estimates, which they term "policy relevant treatment effects," are also useful because they do not require that there be no "defiers" in the language of Angrist et al. (1996). Defiers are individuals who change T in the opposite direction to that intended by the policy, e.g., who have more teen births after the increased availability of contraceptives.

such events may greatly restrict the set of research questions that can be studied. It would not be useful for scientific advance if questions where no instrument is available were simply left unstudied.

A variety of approaches are possible in this case. One is simply to apply OLS to the (y, T, X) relationship and to make a priori arguments on the degree of bias expected. These arguments will necessarily turn on how well the vector X is capable of capturing the main determinants of y and whether there are likely to be unobservables left out which are correlated with T. The direction and magnitude of bias from any remaining unobservables is often something that can be partially assessed on the basis of intuition and outside evidence. The method of matching, described above, can also be applied to determine whether the functional form of the estimated equation is affecting the conclusions drawn about the causal effect of T on y.

Another approach is to apply more formal sensitivity tests to the model to assess how much the estimated effect of T on y would be affected by different degrees of bias. In the case where all error terms are assumed to be multivariate normal, for example, the bias is captured by a control variable termed the Heckman lambda term (Barnow et al. 1980), and a single parameter – the correlation between the errors in the y equation and the T equation – determines the degree of bias in the coefficient on T. Fixing the correlation coefficient at different values and estimating the model with this restriction can be used to assess how the estimate of the effect of T is affected by the magnitude of the correlation (Robins et al. 2000). At another extreme, one can apply an analysis which determines the maximum degree of bias that might arise, a method of "bounds" analysis most formally developed within economics by Manski (1995). This "worst case" analysis can sometimes show that even in the maximal bias case, the estimated effect of T on y is still of reasonable magnitude on a scientific or policy basis. If the maximal bias results in a reversal of results, however, more restrictions on those bounds are needed to obtain more useful results.

2.6 Summary and Conclusions

Much progress has been made in understanding the estimation and interpretation of causal effects with observational data and how exclusion restrictions, which are an implicit assumption that an experiment exists in nature, can be used to identify those effects. Nevertheless, while a deeper understanding has been achieved, the difficulty of the problem has also become better understood. Most importantly, the criteria for valid and relevant instruments have been shown to be particularly stringent, and the scope of what is learned from instruments which are based on narrow populations is now seen to be possibly quite limited. Assessing the validity of instruments is also particularly problematic, as there are no formal tests for validity in the just-identified case, and resolving why different instruments yield different effect estimates can also be quite challenging. Studying the mechanism by which instruments affect the fraction of the population treated and how that interacts with the mechanism by which treatment affects outcomes is now also recognized as important. Much therefore remains to be done.

References

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

Angrist, J. and G. Imbens (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Society* 90 (June): 431–442.

Angrist, J., G. Imbens and D. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91 (June): 444–472.

Angrist, J. and A. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106 (November): 979–1014.

Angrist, J. and A. Krueger (1999). Empirical Strategies in Labor Economics. In: *Handbook of Labor Economics*, Vol. 3A, eds. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Angrist, J. and A. Krueger (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15 (Fall): 69–85.

Barnow, B., G. Cain and A. Goldberger (1980). Issues in the Analysis of Selectivity Bias. In: *Evaluation Review Studies Annual*, eds. E. Stromsdorfer and G. Farkas. Beverly Hills: Sage.

Björklund, A. and R. Moffitt (1987). The Estimation of Wage and Welfare Gains in Self-Selection Models. *Review of Economics and Statistics* 69: 42–49.

Bound, J. and G. Solon (1999). Double Trouble: On the Value of Twins-Based Estimation of the Return to Schooling. *Economics of Education Review* 18 (April): 169–182.

Cameron, A.C. and P. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica* 69 (September): 1127–1160.

Carneiro, P.J., J. Heckman and E. Vytlacil (2006). *Estimating Marginal and Average Returns to Education*. New York: Mimeo.

Chevalier, A. and T. Viitanen (2003). The Long-Run Labour Market Consequences of Teenage Motherhood in Britain. *Journal of Population Economics* 16: 323–343.

Currie, J. and J. Gruber (1996). Health Insurance Eligibility, Utilization of Medical Care, and Child Health. *Quarterly Journal of Economics* 111 (May): 431–466.

Geronimus, A. and S. Korenman (1992). The Socioeconomic Consequences of Teen Childbearing Reconsidered. *Quarterly Journal of Economics* 107 (November): 1187–1214.

Heckman, J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 46: 931–960.

Heckman, J. and R. Robb (1985). Alternative Methods for Evaluating the Impact of Interventions. In: *Longitudinal Analysis of Labor Market Data*, eds. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Heckman, J., S. Urzua and E. Vytlacil (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics* 88 (August): 389–432.

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

Heckman, J. and E. Vytlacil (2001). Policy-Relevant Treatment Effects. *American Economic Review* 91 (May): 107–111.

Heckman, J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* 73 (May): 669–738.

Hoffman, S., E.M. Foster and F. Furstenberg (1993). Re-evaluating the Costs of Teenage Childbearing. *Demography* 30 (February): 1–13.

Hotz, V.J., S. McElroy and S. Sanders (1997a). The Impacts of Teenage Childbearing on the Mothers and the Consequences of those Impacts for the Government. In: *Kids Having Kids: Economic Costs and Social Consequences of Teen Pregnancy*, ed. R. Maynard. Washington: Urban Institute Press.

Hotz, V.J., C. Mullin and S. Sanders (1997b). Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effect of Teenage Childbearing. *Review of Economic Studies* 64 (October): 575–603.

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics* 86 (February): 4–29.

Imbens, G. and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62: 467–76.

Imbens, G. and T. Lemieux (2008). Regression Continuity Designs: A Guide for Practice. *Journal of Econometrics* 142 (February): 615–35.

Klepinger, D., S. Lundberg and R. Plotnick (1999). How Does Adolescent Fertility Affect the Human Capital and Wages of Young Women? *Journal of Human Resources* 34 (Summer): 421–448.

Lee, L.F. (1979). Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables. *Econometrica* 47: 977–996.

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

Moffitt, R. (2003). Causal Analysis in Population Research: An Economist's Perspective. *Population and Development Review* 29 (September): 448–458.

Moffitt, R. (2005). Remarks on the Analysis of Causal Relationships in Population Research. *Demography* 42 (February): 91–108.

Moffitt, R. (Forthcoming). Estimating Marginal Treatment Effects in Heterogeneous Populations. *Annales d'Economie et de Statistique*.

Moffitt, R. and M. Ver Ploeg (eds.) (2001). *Evaluating Welfare Reform in an Era of Transition*. Washington: National Research Council.

Reinhold, S. (2007). *Essays in Demographic Economics*. Unpublished Ph.D. dissertation, Johns Hopkins University.

Ribar, D. (1994). Teenage Fertility and High School Completion. *Review of Economics and Statistics* 76 (August): 413–424.

Ribar, D. (1999). The Socioeconomic Consequences of Young Women's Childbearing: Reconciling Disparate Evidence. *Journal of Population Economics* 12 (November): 547–565.

Robins, J., A. Rotnitzky and D. Scharfstein (2000). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, eds. E.M. Halloran and D. Berry. New York: Springer-Verlag.

Rosenzweig, M. and K. Wolpin (2000). Natural 'Natural Experiments' in Economics. *Journal of Economic Literature* 38 (December): 827–874.

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

Stock, J., J. Wright and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics* 20 (October): 518–529.

Stock, J. and M. Yogo (2005). Testing for Weak Instruments in Linear IV Regression. In: *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds. D. Andrews and J. Stock. New York: Cambridge University Press.