Causal Analysis in Population Research: An Economist's Perspective

Robert Moffitt

Population and Development Review, Vol. 29, No. 3. (Sep., 2003), pp. 448-458.

Stable URL:
http://links.jstor.org/sici?sici=0098-7921%28200309%2929%3A3%3C448%3ACAIPRA%3E2.0.CO%3B2-K

Population and Development Review is currently published by Population Council.

Causal Analysis in Population Research: An Economist's Perspective

ROBERT MOFFITT

MORE THAN MANY other social science disciplines, population studies has a long history of noncausal descriptive analysis. The estimation of aggregate vital rates, the construction and estimation of life tables, and the description of population dynamics are just three examples. The long attention paid to these kinds of topics has led to sophistication in technique and constitutes a major contribution to social science knowledge. In recent years, however, population studies has developed an interest in many other issues that have an explicit causal dimension.

This essay provides an interpretation of the issues, though not one that all economists will necessarily agree with. There is by no means a settled and accepted set of principles for addressing causal questions in economics or even for the proper relative role of causal and noncausal research, for there are significant differences within the discipline in viewpoint, both philosophically and at the practical level.

The economic framework for causal analysis

The modern formulation of the problem of causal analysis is based on the fundamental notion of a counterfactual for an individual, state, country, or other unit (henceforth, "individual" is used for convenience even though the unit of analysis can be anything). Every individual has two possible, or potential, outcomes, Y_1 and Y_0, where Y_1 is the outcome if the individual experiences a particular event or takes a particular action, and Y_0 is the outcome if he or she does not experience it, all else held fixed. For example, Y_1 might be the birthweight of a child if the mother has smoked during pregnancy and Y_0 the birthweight if the mother has not smoked during pregnancy. The difference between the two is the causal effect of the event or action on Y for that individual. Only one of the outcomes is observed in the data because an individual cannot do two things at the same time. The other unobserved outcome—the one not chosen—is called the counterfactual for that individual.

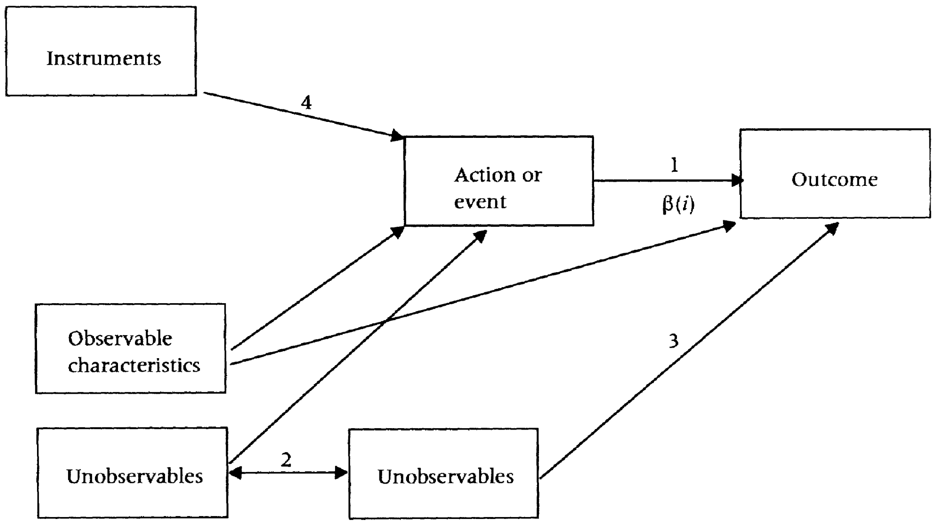
This model is associated with the work of Rubin (1974) in statistics and Heckman (1978) in economics, but also is traceable to the conditional logit model of McFadden (1974) and even the economic model of simultaneous equations developed in the 1940s, which was implicitly a model of counterfactuals. In addition, even the simple linear regression model, when interpreted as a causal relationship, is implicitly a model of counterfactuals if the coefficient on X is interpreted as the difference that would occur in Y if a particular individual were exogenously given one more unit of X.

A basic point in this literature is that the causal effect of the event on Y cannot be estimated without some type of minimal assumption or restriction, even in principle, because of the inherent unobservability of the counterfactual, and that such minimal assumptions and restrictions cannot be formally tested. Consequently, they must be justified or rationalized on the basis of a priori argument, outside evidence, intuition, theory, or some other informal means. This is partly the basis for the statement by Heckman (2000) that "there is no mechanical algorithm for producing a set of 'assumption free' facts or causal estimates based on those facts."

Most of the important causal questions in population research and economic research arise when the variable in question whose effect on Y is the object of interest is potentially endogenous, which is equivalent to saying that those individuals with differing values of the variable may differ in unobserved ways that make an OLS regression coefficient an invalid estimate of the causal effect for an individual. In economics, while it is understood that a wide variety of alternative restrictions are available (in fact, an infinite number), the most common method in practice is to apply what are known as exclusion restrictions, which are sometimes also called instruments.

Figure 1 illustrates the modern causal model with instruments. The object of interest in the figure is arrow 1, denoted $\beta(i)$, which is the true effect of an action or event on an outcome (Y) for an individual i. For example, the action or event may be whether the mother smokes, and the outcome might be child birthweight. Outcomes are affected by observable characteristics and by many things we cannot measure (unobservables). The problem of endogeneity arises because the action-event is also affected by things we cannot measure, for we can never fully explain and measure the determinants of why some individuals take an action or experience an event and others do not (e.g., why some mothers smoke and others do not). If those unobservables are correlated with the unobservables affecting outcomes, a spurious correlation between actions-events and outcomes is set up, working through channels 2 and 3 in the figure. So, for example, if

FIGURE 1 The modern causal model with instruments



women who smoke are disproportionately those with disadvantaged backgrounds in ways we cannot measure, and if disadvantaged background factors also lead to low birthweight, we will observe a negative correlation between smoking and birthweight not because smoking itself causes low birthweight but because women who smoke would have had low birthweights anyway, even if they had not smoked.

The solution based on exclusion restrictions, or instruments, is denoted in the figure by variables ("instruments") that affect the action or event and therefore indirectly affect outcomes. The "exclusion" assumption is that those instruments do not directly affect outcomes, nor are they correlated with the unobservables affecting outcomes. Therefore, if the instruments are observed to be correlated with outcomes, this must mean that $\beta(i)$ is nonzero because it is assumed that the *only* possible channel of effect by which the instruments could affect outcomes is through the action or event. For the birthweight example, the price of cigarettes might be a valid instrument, assuming that the price varies cross-sectionally. It should affect the likelihood of being a smoker, but there is no reason to expect it to directly affect birthweight through any other channel. The choice of instruments, and the question of whether any valid instrument is available at all, are the critical issues in the estimation of causal relationships.

The presumption that the causal effect of interest is different for different individuals is an important aspect of the model because it implies that only the average effect for "switchers" can be estimated. For example, suppose that the variation in cigarette prices in the data induces a variation in the fraction smoking from a minimum of 30 percent of the population to a maximum of 40 percent. The price variation allows the estimation of the average β for the 10 percent of the population who were affected by that variation (the "switchers"). What cannot be estimated is the average β in the entire population because that would require having an instrument which moved the fraction of smokers from 0 percent to 100 percent, thereby permitting the researcher to observe how birthweight changes as the entire population goes from not smoking to smoking (Imbens and Angrist 1994; Heckman and Vytlacil 1999, 2001).

Sometimes it is of sufficient interest to learn only the total, or net, effect of the instrument on outcomes, without learning how that effect works through the action or event of interest. For example, it may be sufficient for the government to know how making contraceptives more readily available affects pregnancy and fertility outcomes on net, with the presumption that the mechanism by which that effect occurs is contraceptive use but without needing actual knowledge of, or data on, the extent to which contraceptive use has changed. This is called a "reduced form" and is just the effect of the instrument on the outcome. The danger in reduced-form estimation lies precisely in the fact that the channel by which the instrument affects outcomes is not estimated. Nevertheless, some economists maintain that reduced-form estimation is sufficient if the only aim of the researcher is to know the effect of the instrument on outcomes and it does not matter what the channel of effect is (e.g., if the government just wants to know the effects of its policies and does not care why they arise).

Types of exclusions

While there cannot be any hard rule on what types of variables are appropriate, especially given the breadth of possible applications, it is possible to mention four types that have been used in a number of different applications: environmental or ecological variables, demographic group variables, twin and sibling relations, and natural experiments.

Environmental or ecological variables measure some aspect of the geographic area in which an individual resides. The unemployment rate, average wage rate, average price of child care, cigarette prices, and availability of contraceptives in an area are examples, as are government policies in the form of laws, benefit levels, tax rates, and regulations that cover an area. The use of these variables is based on the assumption that the individual has little effect, or control, over characteristics of the area in which he is located, and that the variables are thus at least one step removed from the individual's own personal characteristics or actions; they are likely to be exogenous to any given individual's outcomes but should also affect the individual action or event of interest. Sometimes the instrument is taken to be the change in the environmental variable from one time to another in the area in which the individual resides, as in the so-called area fixed effects model, which partly addresses the danger of ecological fallacies.[1]

The objections to this type of instrument are that it requires the assumption that residential location is exogenous, that the ecological correlation problem may still be present even in the area fixed effects model if there are unobservables that are changing over time, that the area fixed effects model assumes that the response to a change in policy is immediate and permanent with no lagged adjustment, that ecological variables are too many steps removed from individual actions and events, and that too few other area characteristics are generally controlled for.

Demographic, or population-segment characteristics fixed effects models (sometimes called "difference in difference" models) assume that nationwide policies affect different individuals differently (e.g., married vs. unmarried women, single mothers vs. single childless women, or even men vs. women) and that the differences in their changes in outcomes over time measure the differences in policies. This requires the assumption that the outcomes of each of the demographic groups would change over time at the same rate in the absence of a change in policy. The method has many of the strengths and weaknesses of the geographic approach, but it has the additional weakness of having to assume that demographic structure is exogenous with respect to changes in policy and that the outcomes of different demographic groups obey the constant growth rate in the absence of policy change just noted.

Twin and sibling models use the within-family variation in the action or event as the implicit instrument. Although twin and sibling models have a long tradition in the estimation of nature–nurture models, where the common family effect is assumed to be genetic, they are more controversial when the common family effect arises from certain socioeconomic variables that have developed over time. In that case there is less assurance that twin and sibling differences are themselves exogenous (see, e.g., Bound and Solon 1999). Also, using twin and sibling differences without any theoretical framework can lead to misspecification of the rest of the equation and hence to incorrect inferences (Rosenzweig and Wolpin 2000).

Natural experiment methods constitute a collection of approaches that seek variables which measure events that have arisen from random, possibly sudden, and usually unpredictable changes in environmental or personal variables, almost always in rather narrow segments of the population. For example, a law might be passed that affects welfare benefits for children aged 4–6 years in one state but not in an adjacent state with a similar economic environment and social composition; the 4–6-year-olds in the two states are compared over time to assess the effect of the law on various child outcomes. Miscarriage, which affects the timing of fertility and has been used to estimate the effect of fertility on outcomes, is another example (see below). Another instance is month of birth, which affects the age at which children are able to enter school according to government rules, which in turn affects completed years of education and can be used to examine the effect of education on various outcomes (see Duncan et al. 2003 for examples in the field of child development).

The analogy to randomized trials, which is often made for all the types of instruments discussed here, is often useful but has pitfalls. One is that it tends to focus attention on the feasibility of randomization, when all that is needed for the validity of an instrument is exogeneity, regardless of whether an experiment would be feasible. Another pitfall is that randomized trials are necessarily reduced-form estimations, for the experimental-control dummy is always the instrument, not the action or event itself, because the latter is a choice variable that cannot be forced on an individual. This means that randomized trials can never estimate channels for the effects of the treatment.

An objection sometimes made to natural experiment instruments is that they are not valid (e.g., they are not truly excludable nor truly exogenous). However, the pure examples of the natural experiment approach that use valid instruments illustrate a tradeoff between what the literature on randomized trials calls internal and external validity. Internal validity holds when the treatment estimate in an experiment is unbiased for the population considered in the experiment and for the particular policy or program tested; in other words, it holds when the experiment was actually carried out correctly with a valid randomization and with no contamination between the experimental and control groups, no biasing attrition, or related problems. External validity holds when the estimate generalizes to a larger population than that in the experiment, or to a larger set of treatments (policies) of interest. The natural experiment examples just mentioned represent an extreme attempt to maximize internal validity (i.e., exogeneity), although they typically result in significant loss in external validity (generalizability). Variation in rules in only two states that apply only to 4–6-year-olds may not generalize to other states or other ages; the fertility variation induced by miscarriage may be small in magnitude and may not generalize to fertility variations caused by other forces or to women with low propensities to miscarry; variations in educational attainment induced solely by month of birth may not generalize to variations induced by changes in educational policy; and so on. The loss in external validity that arises in the natural experiment approach is particularly troublesome for population research, which has a strong tradition of working at the population level and of establishing population-level relationships.

The issue of external validity is often also raised in discussions of twin and sibling models; these always necessarily require subsampling to the population of families with twins or siblings, which may be quite different from the larger population. In addition, twin and sibling models require an additional subsampling down to families in which the twins or siblings have different values of the action or event (e.g., where at least one sibling has a teenage birth and another does not; one graduates from college and another does not; and so on).

The problem of external validity is closely related to the problem of extrapolation, which means the applicability of the results to data points outside the sample actually used in the analysis. Any instrument necessarily represents only one type of policy, environment, demographic type, or other force. Most theoretical models imply, to the contrary, that there exists an effect, $\beta(i)$, which is *the* effect of the action or event, and is independent of the particular cause that induces it to change. Yet the approach to causal modeling discussed in this essay cannot guarantee that any such true effect exists, because the only type of effect estimated is the effect of variations in the action or event induced by a particular instrument. Whether abstract effects exist that are independent of the particular causes that induce them is a difficult question not easily resolved.

In addition, a narrower issue of extrapolation necessarily arises for the reason noted earlier, namely that any instrument only induces variation in the action or event within a particular range (e.g., from 30 percent to 40 percent in the smoking example). Extrapolation to the rest of the population requires additional assumptions. Other things equal, therefore, instruments that induce greater variation in the action or event are to be preferred to those that induce lesser variation. Unfortunately, typically there is a tradeoff between the strength of the argument for internal validity and the range of variation in the action or event induced; this is clearest in the natural experiment approach, which often generates only small variation. The issue of extrapolation is also raised in reduced-form estimation because such estimation does not identify the channel of effect, and knowledge of channels can be important in extrapolation, even if only informally.

An illustration: The effect of teenage childbearing on child outcomes

There is a significant literature in the United States on the causes and consequences of teenage childbearing, particularly childbearing out of wedlock. Teenage childbearing increased over the 1970s and 1980s, and it is feared that having children too early may hurt the educational, income, and possibly marital prospects of the mother. These negative consequences could, in turn, disadvantage the children born to such mothers. One of the typical outcomes for a child is his or her cognitive ability at, say, age 5. The counterfactual in this case is the cognitive ability of a child at age 5 if his mother had delayed having him until she was older. The selectivity issue is that a simple comparison of cognitive outcomes of children born to teenage mothers and to older mothers may be biased because of unobserved differences between them.

An issue for this counterfactual is what else is held constant when imagining birth postponement, and therefore to what extent the channel of effect is being captured. The problem is that there are many different channels. Postponement may lead to different outcomes for education, earnings, family income, marital status, and family structure. Postponement may also simply lead to greater maturity, which could have effects on parenting ability. Most studies that attempt to address the endogeneity problem have not been able to determine the mechanism by which postponement affects outcomes, which makes it difficult to interpret the results.

The multiplicity of channels is closely tied to the issue of whether all causes of postponement would have the same effect. If the cause of postponement is that more resources are put into local schools, improving their quality and making educational continuation more attractive, that will clearly affect educational attainment. Because education has its own effects on an individual's breadth of perspective, it could affect childrearing. If the cause of the postponement is a sudden increase in the number of jobs available in the local labor market, women may drop out of school to take jobs and postpone births because they are able to work. This is likely to have different effects on women's subsequent life experiences. Another cause of postponement could be an improvement in contraceptive technology. In this case, women could engage in sexual activity without pregnancy, but whether that would lead to additional educational or employment outcomes depends greatly on the attractiveness of those two alternative uses of time, and on how many women would avail themselves of it. If the cause of postponement is a reduction in the supply of available sexual partners, that would have obviously different effects on future marriage rates than if the cause is one of the others mentioned. If the cause of postponement is a national or local media campaign to reduce teen pregnancy, the effects could again depend on what the affected women would choose to do with their lives in the absence of having children while young. It would seem, therefore, that the simple question of *the* effect of birth postponement on child outcomes is ill-posed, and needs a theoretical framework within which the reason for postponement is part of the model and is tied to specific channels.

The search for a candidate instrument involves a search for a variable that induces a birth postponement for reasons not related to child outcomes. Ecological or environmental variables that may affect birth postponement are many: labor market opportunities, quality of education in the area, and the sex ratio are three candidates tied to specific causes mentioned above. However, these three are unlikely to satisfy the exclusion requirement because they are likely to affect outcomes in more direct ways. Two ecological variables that have been argued to satisfy this criterion are the local availability of contraception and a state's abortion po cy. Both may affect teenage birth rates and, at least at first blush, do not appear to affect child outcomes directly. As with all ecological variables, there is some danger that states and areas with differing contraceptive availability and abortion policies differ in some other unobserved way, leading to spuriously estimated effects. Examining changes in child outcomes over time in response to changes in contraceptive availability or changes in abortion policy must confront the issue of time-varying unobservables and lagged responses. Changes in contraceptive availability and abortion policy may also have modest effects on the rate of teenage childbearing, with the consequence that the results do not generalize to what those effects would be for larger segments of the population.

Sibling and twin models have been used as well (Geronimus and Korenman 1992; Hoffman et al. 1993; Grogger and Bronars 1993), albeit for a case where the outcome for the mother was the object of interest rather than that for the child. The strengths and weaknesses of the approach for this application are the same as those mentioned in general above. The analysis must be conducted on a relatively small subset of the population—families with at least two siblings, one of whom has had a teenage birth and one of whom has had a later birth—and there must be no intrafamily, individual differences between the two women that led them to make different birth timing choices that might be correlated with the outcomes for their children. The latter assumption seems particularly strong.

A natural experiment instrument that has been used (though, again, for adult rather than child outcomes) is the occurrence of miscarriage (Hotz et al. 1997). If miscarriage is random, reduced forms yield valid estimates of the effect of miscarriage; if miscarriage is also excluded from the Y equation, then the mechanism—birth postponement—has been correctly identified. Whether or not miscarriage is random depends on whether it is a result of biological or behavioral or environmental factors or, even if the first, whether those are correlated with behavioral or environmental factors (which are likely to be correlated directly with maternal and child outcomes). The excludability of miscarriage from the main equation requires the assumption that having a miscarriage has no direct effects, either economic or affective, on the mother or the later-born child. Questions of external validity would arise if the population of mothers with miscarriages is a specialized subset of the population, and if the change in percent of teen births resulting from a particular miscarriage rate is small.

Conclusions

This review of the analysis of causal relationships illustrates the difficulties in reaching conclusions about those relationships. There are very few cases where a single clearly superior method of addressing the traditional self-selection and endogeneity problem can be found, and most of the methods used in the past are open to serious objections. This situation has led to a rather pessimistic view of the progress of the field and of the prospects for increasing knowledge in the future, and it signals a retreat both in what is known and even in what can be known. Indeed, it may be that certain types of causal effects are essentially unknowable with any reasonable level of confidence. Modesty of claims for truth is the clear lesson from this review.

The direction in which it is best to proceed in light of these difficulties is the question that will be addressed by the discipline in the future. To be sure, the solution is not to return to ignoring the problem altogether. Indeed, given that average practice at the moment often still ignores the endogeneity problem altogether, or addresses it in ways that are now thought to be inadequate, the direction in which movement should be made is to raise the standards for acceptable solutions and to be more discriminating and careful in their use. More attention to theory, to mechanisms, and to the possible invalidity of particular exclusion restrictions is needed in order to change the direction of research toward a more cautious and considered approach.

At the same time, there is a danger in maximizing internal validity at the expense of external validity. To do so would lead to a field consisting only of very narrowly defined exercises without generalizability and to a collection of miscellaneous facts that do not add up to any general knowledge. Such an approach is intellectually constricting and does not allow the pursuit of the larger social science questions and the explanation of societal-wide trends that motivate much of social science research.

To achieve the correct balance requires a weighing of evidence from different studies with different strengths and weaknesses to achieve consensus. To be sure, it is fair to rule some types of evidence out of court completely if the problem of controlling for endogeneity is too severe, but there still remains a larger set of studies with different mixes of internal and external validity that should be given some positive weight when drawing conclusions on what the best estimate is. This approach calls for a wide variety of evidence and for more research that attempts to reconcile the different findings of different studies and attempts syntheses in literatures where different approaches are taken. Areas where a variety of approaches yield similar results are those where synthesis should be easy, while those where that variety yields dissimilar results are the most difficult. Formal theory as well as informal evidence from ethnographic and other accounts should also be given positive weight. Such a programmatic, considered, and synthetic approach to research questions is likely to yield the most progress in light of the impossibility that any single study or set of studies can provide the best answer.

Notes

A previous version of this note was presented at the Annual Meeting of the Population Association of America, May 2003, Minneapolis. The author thanks Greg Duncan and Geoffrey McNicoll for comments.

1 An ecological fallacy occurs when the characteristics of the area in which an individual resides are correlated with that individual's own outcomes and actions, but only coincidentally and not because of a true causal connection.

References

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.

Duncan, G., K. Magnuson, and J. Ludwig. 2003. "The endogeneity problem in developmental studies," mimeographed, Northwestern University.

Geronimus, A. and S. Korenman. 1992. "The socioeconomic consequences of teen childbearing reconsidered," *Quarterly Journal of Economics* 107 (November): 1187–1214.

Grogger, J. and S. Bronars. 1993. "The socioeconomic consequences of teen childbearing: Findings from a natural experiment," *Family Planning Perspectives* 25 (July–August): 156–161.

Hoffman, S., E. M. Foster, and F. Furstenberg. 1993. "Re-evaluating the costs of teenage childbearing," *Demography* 30 (February): 1–13.

Heckman, J. 1978. "Dummy endogenous variables in a simultaneous equation system," *Econometrica* 46: 931–960.

———. 2000. "Causal parameters and policy analysis in economics: A twentieth century retrospective," *Quarterly Journal of Economics* 115 (February): 45–97.

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.

———. 2001. "Policy-relevant treatment effects," *American Economic Review* 91 (May): 107–111.

Hotz, V. J., S. McElroy, and S. Sanders. 1997. "The impacts of teenage childbearing on the mothers and the consequences of those impacts for the government," in ed. R. Maynard, *Kids Having Kids: Economic Costs and Social Consequences of Teen Pregnancy*. Washington, DC: Urban Institute Press.

Imbens, G. and J. Angrist. 1994. "Identification and estimation of local average treatment effects," *Econometrica* 62: 467–476.

McFadden, D. 1974. "Conditional logit analysis of qualitative choice behavior," in P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press.

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 nonrandomized studies," *Journal of Educational Psychology* 66: 688–701.