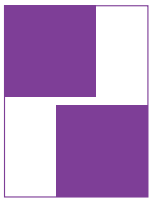


Changes in
HEALTH CARE
FINANCING &
ORGANIZATION

Does X Really Cause Y?

By Bryan Dowd and Robert Town
September 2002

AcademyHealth
Advancing Research, Policy and Practice

AcademyHealth is the national program office for HCFO, an initiative of The Robert Wood Johnson Foundation.

Foreword

Health policy issues often dominate state and federal policymakersÕ agendas. In the most recent session of the United States Congress alone, the House and Senate addressed legislation concerning a patientsÕ bill of rights, prescription drugs for seniors, and generic drug substitution. While politics and legislative realities seem to have starring roles in the process, in most cases, research results and other information can be solid supporting players. Those responsible for therecommendations, if not also thedecisions, strive to increase their knowledge of the problem at hand, as well as the likely impact of the regulation or legislation under consideration. However, they frequently expressfrustration that they do not have objective evidence-based and timely information on which to base their recommendations anddecisions.

Health services researchers, meanwhile, generate information about many of the same pressing health policy issues. Each year millions of dollars are spent by private foundations and the federal government to support health services research designed to produce useful, policy relevant results. A plethora of monthly journals are filled with articles highlighting findings from studies of health care costs, quality, and access, as well as interventions designed to improve health and health care. Many universities and research institutions publish reports with findings of interest to decision-makers.

The challenge for the field, in general, and shared by us at The Robert Wood Johnson FoundationÕs Changes in Health Care Financing and Organization (HCFO) initiative, is to develop effective mechanisms for researchers to make their findings accessible to policymakers seeking information. Typically, with this goal in mind, we present summary research findings in short documents focused on specific issues. Often these summaries clearly delineate the findings, but in an effort to increase their clarity and importance, there is very little explanation of the methods used to develop the findings or caveats that might be applicable in a real-world setting.

This special report, emanating from a HCFO meeting, ÒManaged Care Spillover: Research and Policy Issues,Ó conducted by The Academy for Health Services Research and Health Policy (now AcademyHealth) on November 8, 2001, takes another approach to making research information accessible. We asked Bryan Dowd to assist the audience of senior-level decision-makers in understanding how the findings from studies of the effects of managed care spillover might be used to inform policy discussions. He discussed the challenges to evaluating complex interventions or phenomena using standard econometric techniques. He explained how equations are used to depict the relationships among the key variables being studied, and he identified common sources of bias or error that might affect the results of such analyses. He also pointed out that while research and policy development often require using imperfect information, it is important for those using research findings to ask questions that allow them to identify and account for such imperfections. Participants at the meeting found DowdÕs insights and brief review of econometrics to be helpful and recognized that his guidance would be of general use beyond the specific study findings being disseminated and beyond those participating in the meeting. Therefore, the HCFO program commissioned Dowd to further develop his presentation for broader distribution.

While the report may be somewhat technical for those with no exposure to econometric methods, we hope that it will serve as a useful guide to analysts and information brokers with minimal background in econometric analysis as they seek to fully understand the research findings and reports available. Decision-makers with the time and inclination to go beyond a list of summarized research findings can use this report to ask appropriate questions and provide necessary caveats when considering the validity and applicability of research findings to a specific policy discussion.

Anne K. Gauthier, Program Director
Deborah L. Rogal, Deputy Director

Introduction

Good public policy decisions require reliable information about the causal relationships among variables. Policymakers must understand the way the world works and the likely effects of manipulating the variables that are under their control. The purpose of this paper is to assist policymakers by providing an introduction to some of the problems associated with causal inference from empirical data. We hope that the paper also will be helpful to researchers who are attempting to draw causal inferences from data, or explain their results to policymakers.

Policymakers face a number of problems when presented with results from empirical studies. Sometimes the studies use inappropriate methods or draw unwarranted inferences from the results. Sometimes the results from good studies are subjected to selective interpretation before being passed on to policymakers. We hope that this paper will help researchers produce better research, and help policymakers ask the kinds of questions that will provide the information they need in order to make well-informed decisions.

Our discussion is narrative and intuitive, rather than mathematical. Researchers encountering the subject for the first time can pursue the topics introduced in this paper by picking up any introductory econometrics text and reading the sections on omitted variables bias and simultaneous equations, and then reading the same sections in more advanced texts.

Advanced researchers may find our treatment of the topic frustrating, and even misleading. Only two of many problems associated with causal inference are discussed, and they are discussed only for the case of cross-sectional data, with only brief references to time-series (panel) data[1]. Problems of measurement and of generalizing sample results to inappropriate populations are omitted. The latter problem is assumed to be resolved satisfactorily by the sampling theory of frequentist statistics. Problems of aggregation across subgroups (including SimpsonÕs paradox or the Òecological fallacyÓ and LordÕs paradox) are omitted.

Health services researchers who would like more information on these omitted topics should refer to papers from a conference titled ÒResearch Methodology: Strengthening Causal Interpretations of Nonexperimental Data,Ó sponsored by the (then) Agency for Health Care Policy and Research, in 1990 (AHCPR, 1990).

Background

Does a change in the value of one variable really produce a change in value of another variable? This seemingly simple question has challenged some of the greatest thinkers in history, including Heraclitus, Plato, Aristotle, Galileo, Hobbes, Hume, Kant, and countless other philosophers and scientists. Rigorous treatment of causality in statistical analyses peaked during the last half of the 20th century with the work of the Cowles Commission and its intellectual offspring (Heckman, 1999). Many of the major advances in econometrics during the last four decades have been in the area of causal inference. A number of recent books have been devoted to the subject, including *Causality: Models, Reasoning and Inference* by Judea Pearl (2000), and *Causality and Explanation* by Wesley C. Salmon (1998). The discoveries of quantum mechanics have added to both the mystery and our understanding of causality. At the beginning of the 21st century, the topic of causality is more intriguing and perplexing than at any time in history.

The discoveries of quantum mechanics have added to both the mystery and our understanding of causality. At the beginning of the 21st century, the topic of causality is more intriguing and perplexing than at any time in history.

There are a number of fundamental philosophical issues regarding causality. For example, does causality imply a real, physical connection between causes and effects, or does the appearance of cause and effect exist only in our minds? Can true causal mechanisms ever be established, or is causality merely a probabilistic statement about the likelihood that two events will occur in conjunction? What is a probability, for that matter? How should the fact that many causal theories may be consistent with the same data be incorporated into empirical investigations of causality?

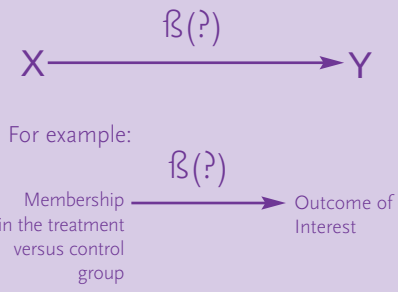
For the most part, this paper sets the philosophical questions aside and focuses on the same topic considered by the Cowles CommissionÑcausality in the context of empirical statistical analyses. An important contribution of the Cowles Commission was the identification of the conditions under which the parameters of a causal model can be recovered from correlational data (Heckman, 1999). The textbook example of this problem is the estimation of demand and supply equations. In theory, market-wide demand curves should represent the negative relationship between output price and demand for the commodity. Market-wide supply curves should represent the positive relationship between output price and supply of the commodity. The Cowles Commission identified the data required to recover a negative coefficient on price in the demand equation and a positive coefficient on price in the supply functions from a single cross-sectional dataset on prices and quantities taken from different market areas.

In health services research, analysts often are concerned with causal effects. Virtually all program evaluation and health outcomes research attempts to establish causal effects. Models of choice, demand or supply relationships, and medical decision-making all have embedded within them the assumption that changing the values of some of the variables will result in changes in the values of other variables. Researchers may soften their rhetorical treatment of causality by saying that the coefficient from a linear regression model is Òthe change in the dependent variable that is associated with a one unit change in X.Ó All too often, however, the analyst would not be offended in the least if the reader assumed that Òif you change X one unit holding the values of the other variables constant, you will get a 10 unit (for example) change in Y.Ó The ceteris paribus condition (i.e., Òholding the effects of other variables constantÓ), may be nothing more than a fond wish, but it often is incorporated unapologetically into the interpretation of the results by elevating the estimated coefficients to the status of partial derivatives.

The Problem

The essential problem is estimating the effect of changing the value of an explanatory variable (X) on the value of a dependent variable (Y). An example would be the effect of moving subjects from the treatment group to the control group in a study of health outcomes, as shown in Figure 1. The estimated causal effect of X on Y is the coefficient §. In a linear regression model, § is the change in Y produced by a one unit change in X. For example, if X is measured in years, § is the effect on Y produced by an additional year of X.

Figure 1:
The Question: What is the Causal Effect of X on Y?



Unfortunately, the real world is more complex than the diagram in Figure 1. Figure 2 shows the empirical problems encountered in the real world. X, Y, W and Z are observed variables and u and v are vectors of unobserved variables, or stochastic error terms. In addition to the causal effect of X on Y, there also may be variables that are causally related to both X and Y (denoted Z and W). If these variables are observed, their effects can be modeled and controlled. If they are not observed, the effect of their omission depends on the way in which they operate. If the omitted variables represent pathways by which X affects Y (i.e., an unobserved W), their effect will be incorporated into the estimated causal effect §. This change in § does not necessarily represent a bias, as long as the analyst recognizes that the estimated effect includes the effect of W.

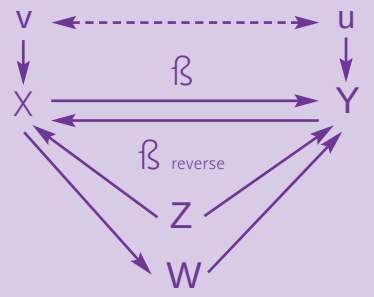
If the omitted variables represent a common cause of X and Y (i.e., an unobserved Z) their omission will result in spurious correlation and biased estimates of the causal effect §. Epidemiologists and sociologists might refer to Z as a confounding variable. Econometricians refer to bias of this type as omitted variables bias

The second problem considered in this paper is reverse causality. Reverse causality means that Y might exert a causal effect on X, in addition to (or instead of) the effect of X on Y. Reverse causality is denoted in Figure 2 by the coefficient labeled ÒßreverseÓ

The essential problem is estimating the effect of changing the value of an explanatory variable (X) on the value of a dependent variable (Y).

Empirical analysis in econometrics usually focuses on the first and second concepts of causality, treating the third concept, antecedence, as a necessary, but not sufficient, condition for causality.

Figure 2:
The Real World



What is to be done about these problems of omitted variables and reverse causality in analyses of empirical data? At first, the outlook seems bleak. Peter Kennedy in his popular *Guide to Econometrics* (1998) says Ò(u)sing the dictionary meaning of causality, it is impossible to test for causality.Ó What is the dictionary meaning of causality? WebsterÕs dictionary (1977) gives the following definition:

Causality: Òperson or thing responsible for an action or result. To make happen; bring about.Ó

This definition is followed by an explanation of five concepts of causality. A health services research example has been added for each one.

1. Cause: A cause helps bring the event about. Example: Increasing the copayment for outpatient office visits will cause a decrease in the demand for visits.
2. Determinant: A determinant fixes the nature of the result. Example: Education is a determinant of health status.
3. Antecedent: An antecedent precedes the result. Example: A rapid increase in health insurance premiums in the early 1990s was an antecedent of the Clinton administrationÕs health care reform proposal.
4. Reason: A reason is a traceable or explainable process relating the cause and result. Example: The fear that they will face high out-of-pocket expenses for uncovered inpatient expenditures is the reason why Medicare beneficiaries buy private supplementary insurance policies.
5. Occasion: An occasion triggers a result and may be different from an underlying cause. Example, the spend-down requirement of Medicaid is a cause of asset shifting from one spouse to another among elderly couples, but a specific adverse health event can be the occasion that triggers the asset shift for a particular couple.

Empirical analysis in econometrics usually focuses on the first and second concepts of causality, treating the third concept, antecedence, as a necessary, but not sufficient, condition for causality. Antecedence has found its way into the econometrics literature as Granger causality, which Maddala (1988) equates to Òprecedence.Ó Granger causality means that past values of X are helpful in predicting the current values of Y, but future values of X are not. Kennedy (1998) notes that econometricians should use the term Granger causality rather than causality, but generally do not. Health services research usually does not focus on Òoccasions,Ó although there certainly are interesting research questions in care-seeking behavior or health plan disenrollment, for example, that focus on the ÒoccasionsÓ that trigger action on the part of the individuals.

The concept of ÒreasonÓ is one that probably could use more attention in health services research. For example, Christianson, et al. (2001) note that many studies examine health outcomes for chronically ill Medicare beneficiaries in HMOs versus FFS Medicare, but only a small number explore the mechanisms that plausibly could produce the observed differences. In his textbook *Causal Analysis*, David Heise (1975) is adamant about the concept of reason, stating that Ò(o)ne event does not directly cause another if no effective operator is available to support the relationship.Ó

A number of formidable barriers stand between the data analyst and the establishment of causality. The first, mentioned earlier, is the ceteris paribus condition that the effects of all other variables must be held constant. In health services research, and particularly in non-experimental settings[3], it often is physically (technically) impossible to hold the effects of all other variables constant, and implausible to argue that they effectively are held constant. The fallback position is to employ theory and statistical modeling to approximate the ideal world of a perfectly controlled experiment.

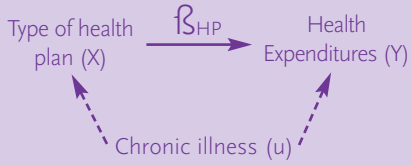
A second threat to causal inference is the fact that any empirically determined relationship among variables is likely to be consistent with a number of different theories. The approach to that problem is to develop a theory, stated in the form of a null hypothesis, and then test whether a dataset (different from the data used to develop the theory) can reject the null hypothesis at some predetermined level of statistical confidence. The result of the test is all one can report at that point. Other hypotheses remain viable until they are rejected, perhaps by developing theory and empirical tests that can distinguish among competing hypotheses.

This paper examines the difficulties associated with testing a particular null hypothesisÑno causal effect of X on YÑin cross-sectional data. It also illustrates the use of theory and statistical modeling to address two specific threats to causal inference: omitted variables and reverse causality.

Omitted variables bias

Suppose we are trying to estimate the causal effect of membership in a particular type of health plan on subsequent health expenditures or utilization of services. This question arises often in comparisons of managed care and fee-for-service health plans in the Medicare program. The problem is shown in Figure 3.

Figure 3:
Spurious Correlation



Chronic illness is unobserved, and is an example of omitted variable bias or “spurious correlation.”

This paper illustrates the use of theory and statistical modeling to address two specific threats to causal inference: omitted variables and reverse causality.

In Figure 3, chronic illness is a variable that is unobservable by the analyst, and thus is an omitted variable in the analysis. Chronic illness is a common cause of both the type of health plan the subject joins (X) and subsequent health expenditures (Y). Because chronic illness is unobserved and cannot be included in the regression equations for Òtype of health planÓ or Òhealth expenditures,Ó it is contained in the unobserved error or disturbance term, denoted Òu.Ó Under the assumption of a linear relationship between health expenditures and types of health plan, we could write the regression equations for health expenditures as:

(Equation 1)

Health Expenditures = β_{HP} Type of Health Plan + u

where u represents chronic illness.

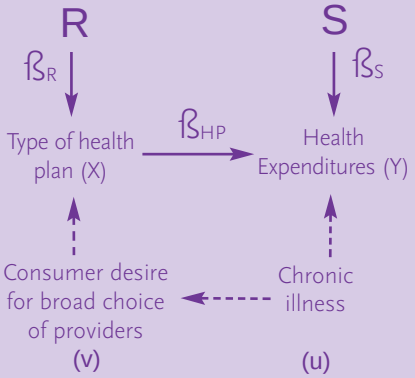
If there are omitted variables like chronic illness in our analysis, then the estimated causal relationship, represented by β_{HP}, will misstate the true causal relationship. For example, we might conclude that expenditures were lower in managed care Medicare+Choice (M+C) plans than in FFS Medicare, but the lower expenditures might be due, in part, to a lower prevalence of chronic illness in the M+C population.

Econometricians say that the estimate of β_{HP} is biased, because the estimated relationship is due, in part, to the effect of chronic illness on both the type of health plan the subjects join and their subsequent health expenditures. The error term ÒuÓ is correlated with the explanatory variable ÒType of Health PlanÓ in the regression equation. Statisticians refer to this type of bias as Òspurious correlation.Ó Epidemiologists would say that chronic illness is an unobserved confounder. In simple models such as the one shown in Figure 3 it is possible to determine the direction of the bias. If the effects of chronic illness on type of health plan and health expenditures have the same sign (e.g., a proportional (positive) or inverse (negative) effect), then the estimate of β_{HP} is too positive. If the effects of chronic illness on type of health plan and health expenditures have opposite signs, the estimate of β_{HP} is too negative. In more complex models, it often is difficult to determine the direction of the bias, because the direction depends on the correlations among all the variables, both observed and unobserved, in the model.

Figure 4 shows another version of omitted variables bias. Here, the problem is not a common omitted variable, but correlation among the omitted variables that cause the consumer to join one type of health plan versus another and the omitted variables that cause health expenditures. The result, however, is the same as in Figure 3Ña biased estimate of the causal effect of X on Y, represented by β_{HP}.

In more complex models, it often is difficult to determine the direction of the bias, because the direction depends on the correlations among all the variables, both observed and unobserved, in the model.

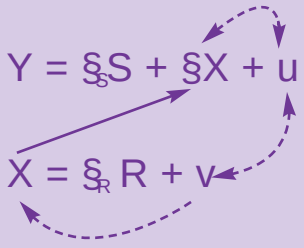
Figure 4:
Spurious Correlation



Econometricians refer to the problem illustrated in Figures 3 and 4 as Ò2correlated errors across equations.Ó What they mean is the unobserved error term in the equation describing how Òtype of health planÓ takes on its value (i.e., Òconsumer desire for broad choice of providersÓ) is correlated with the error term in the equation describing how Òhealth expendituresÓ takes on its value (i.e., Òchronic illnessÓ).

Figure 5 shows how the biased estimate of § arises. The unobserved determinants of X (type of health plan), denoted v, cause changes in X. However, X also appears in the Y (health expenditure) equation. That alone would not be problematic, but the correlation of u and v means that X and u are correlated in the Y equation, which leads to biased estimates of §, just as in Equation 1.[5]

Figure 5:
How the Correlation of X and u Arises



While the problem of omitted variables can be acute, it often causes more consternation than necessary. It is important to understand the limited nature of the cases in which omitted variables result in biased estimates of the causal effect of X on Y from ordinary least squares (OLS) regression.[6] All stochastic regression equations have omitted or unobserved variables, which is why they have a stochastic error term. All unobserved variables reduce the explained variation in the dependent variable, which, in turn, reduces the statistical power of the analysis. But not all omitted variables result in biased estimates of causal effects. Only variables that causally affect both the dependent variable (Y) and an explanatory variable (X) result in biased estimates of §.

Figure 6 shows a case in which the omitted variable is correlated with both type of health plan and health expenditures, but estimates of the causal effect ($§_{HP}$) are not necessarily biased. The model in Figure 6 appears, at first, to be similar to the models in Figures 3 and 4 because the variable Òphysician practice styleÓ is unobserved and is correlated with both X (type of health plan) and Y (health expenditures). However, physician practice style is not a source of spurious correlation, because in the version of the theory shown in the diagram, it does not cause the consumer to join one health plan versus another. Rather, it is a product of the hiring practices of the health plan. It is one of the ways in which health plan membership affects health expenditures. If the physiciansÕ practice style is not observed, the effect of physician practice style will be incorporated into the estimated coefficient $§_{HP}$. $§_{HP}$ is not biased, as long as the analyst understands that it represents the

While the problem of omitted variables can be acute, it often causes more consternation than necessary.

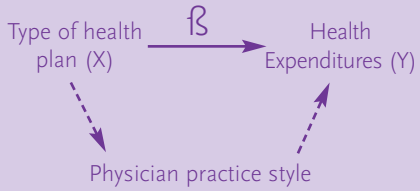
ÒfullÓ effect of type of health plan on expenditures, rather than the ÒpartialÓ or ÒresidualÓ effect, controlling for physician practice style.

In some analyses, the objective is to understand all the ways in which X might affect Y. Data might be collected on variables such as physician practice style and added to the model with the hope that the estimated coefficient β_{HP} (i.e., the residual effect controlling for all observed pathway variables) will become zero.

Another way to distinguish the examples in Figures 3 and 4 from the example in Figure 6 is to note that Òtype of health planÓ in Figures 3 and 4 is an endogenous stochastic variable. That means that Òtype of health planÓ is itself a dependent variable in an equation that contains the stochastic error term v. However, in Figure 6, Òtype of health planÓ is predetermined relative to the value of health expenditures that it helps to produce. It is not caused by anything in the model. Rather than being an endogenous stochastic variable, Òtype of health planÓ in Figure 6 is said to be fixed in repeated samples.

Policymakers and researchers often are interested in establishing the causal consequences of past policy initiatives.

Figure 6:
Unobserved Pathways



The issues raised in Figures 3, 4, and 6 illustrate a problem that has received considerable attention in the econometrics literature. Policymakers and researchers often are interested in establishing the causal consequences of past policy initiatives. Typically, the causal inference must be based on data from past initiatives. The problem is that past causal relationships, even if established with considerable certainty, need not imply future causal relationships. In the economics literature, this issue is referred to as the Lucas Critique after Robert LucasÕs work on the relationship between expectations, monetary policy, inflation, and unemployment.

A simple example illustrates the problem. Suppose that at some point in the past, Medicaid eligibility was expanded in some sites, and a subsequent improvement in the health status of the newly eligible population was observed. Suppose further that a carefully designed research project determined with reasonable certainty that there was, in fact, a causal link between the expansion of eligibility and the improvement in health status. Finally, suppose that Medicaid expansion has an impact on health status through the following causal pathway:

a) individuals who previously were uninsured sign-up for the newly expanded Medicaid insurance;
b) the increase in insurance coverage leads to demand for more health care services;
c) providers are willing to supply services to the newly entitled beneficiaries; and finally
d) the additional health care consumption leads to an improvement in health.

Will further expansions in the same sites result in an even greater improvement in health status? Will expansion of eligibility in new sites produce the same results as in the original sites? Unfortunately, the original study does not provide clear answers to these questions.

The first potential problem is a product of the original research design. When research funds are limited, government agencies, foundations, and researchers may legitimately choose to test a new intervention in sites where it has the greatest chance of success. The theory is that if the intervention fails there, it would fail everywhere. However, an intervention that succeeds in the most favorable site could fail everywhere else. Thus, former success is not a good indicator of future success. This is a case of omitted variables bias, corresponding to Figures 3 and 4. The omitted variables are characteristics of the site that influenced which sites received the intervention (X), and the likelihood of the interventionÕs success in those sites (Y). This is what econometricians call sample selection bias.

The second potential problem is that the causal effect, though real in the original sites, may have been exhausted at the level of Medicaid expansion that was implemented in the sites, and further expansion may have no further effect on health status. Incorrect inference could arise from failure to realize that the relationship between Medicaid expansion and health status was non-linear. This failure could be depicted as a missing pathway, as in Figure 6. Suppose that the level of expansion is inversely related to the severity of illness of individuals in the program, and the severity of illness, in turn, is positively related to improvement in health (e.g., greater severity implies greater improvement). The misinterpretation of the results from the original data is a failure to recognize that the estimated § relating Medicaid expansion to health status in the original study was the full effect, rather than the partial effect, controlling for the severity of illness among newly eligible enrollees.

There is no easy solution to these problems. In particular, there is no easy solution that relies solely on econometric techniques as opposed to a more thorough substantive understanding of the phenomena that produced the observed data, and there is no consensus in the economics profession on the best approach. Some economists advocate that policy analysis should be based only on the estimation of ÒdeepÓ or ÒstructuralÓ parameters. That is, the econometric model should include and estimate the causal pathways by which a policy has its effects on outcomes of interest. Unfortunately, while this approach has theoretical appeal, these Òstructural modelsÓ often require many untenable assumptions. Other economists advocate a reduced form approach, in which the effects of the policy are summarized in a single parameter estimate, and the researcherÕs focus is on controlling for potential confounders.

Some economists advocate that policy analysis should be based only on the estimation of “deep” or “structural” parameters. That is, the econometric model should include and estimate the causal pathways by which a policy has its effects on outcomes of interest.

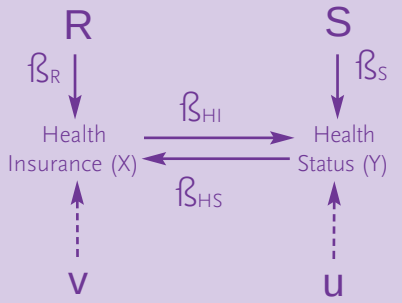
Reverse causality

To continue an example from the previous section, suppose we are interested in estimating the effect of health insurance on the health status of a population. As shown in Figure 7, health insurance indeed might affect health status by providing increased access to health care

Reverse causality can be present in addition to the causal effect that we are trying to detect, or it can be present instead of the effect we are trying to detect.

services, but consumers in poor health status might find it difficult to purchase health insurance in the individual commercial health insurance market. (They almost certainly will find it expensive.) In addition, poor health also might affect the consumerÕs access to employment, and thus to employment-based health insurance. So health insurance might affect health status and health status might affect health insurance, as well. This is an example of reverse causality. Reverse causality can be present in addition to the causal effect that we are trying to detect, or it can be present instead of the effect we are trying to detect.

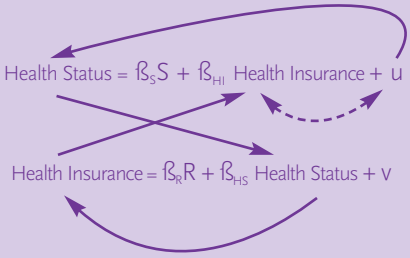
Figure 7:
Reverse Causality



In one sense, the problem is similar to omitted variables bias in that health insurance (X) is a stochastic explanatory variable, requiring its own equation, and determined, in part, by the stochastic error term v. However, the problem with reverse causality is different from omitted variables in that u and v do not need to be correlated in order for bias to arise.

Remembering that bias arises due to a correlation of X and u in the Y equation, Figure 8 shows how that correlation arises in the case of reverse causality. There is some unobserved variable u (e.g., prior onset of chronic illness) that affects health status (Y). Health status, however, is a causal determinant of the consumerÕs ability to obtain health insurance (X) as shown in the second equation. This causal effect is distinct from the causal effect of health insurance on health status. But health insurance appears in the health status equation, and thus health insurance and u are correlated (as shown by the dotted arrow). The error terms u and v might or might not be correlated, but their correlation is not necessary to produce biased estimates of β_{HI}. In a simple linear regression equation, the estimated causal effect of health insurance on health status is due, in part, to the causal effect of health status on health insurance.

Figure 8:
How Correlation of X and U Arises in the Case of Reverse Causality



Although the problems of omitted variables bias and reverse causality are substantively different, they result in the same source of biasÑcorrelation of the explanatory variable or variables in a regression equation with the error term in the equation. The next section discusses approaches to correcting the problem.

How researchers approach these problems

The problems of omitted variables and reverse causality appear formidable, and in many cases they are difficult or impossible to resolve. However, analytic approaches have been developed, and it is helpful for policymakers to have an intuitive understanding of those approaches. Some of the approaches are essentially different, while others simply are variations on a theme. In some cases, the same approach is applicable to both omitted variables and reverse causality, while other approaches apply only to one problem or the other.

Omitted variables

As discussed in the previous section, the special problem of omitted variables bias arises not from unobserved variables per se, but unobserved variables that cause both X and Y, as shown in Figure 3, or the correlation of unobserved variables that cause both X and Y, as shown in Figures 4 and 5. Remember that the essential problem is the correlation of X with the error term u in the Y equation.

There are three approaches to the problem:

1. Collect additional data on the unobserved variables and add them to the analysis.
2. Attempt to manipulate X in a way that has no effect on Y other than through the induced changes in the values of X.
3. Model the correlation of the error terms in the X and Y equation as part of the estimation approach.

Collecting additional data

Although the value of collecting additional data never should be underestimated, there can be important costs in both time and money associated with additional data collection, and it may not ever be possible for the analyst to feel confident that all the unobserved variables that could contribute to omitted variables bias have been collected. For that reason, the possibility of collecting additional data is acknowledged, but the remainder of the discussion focuses on the second and third approaches to the problem.

Although the value of collecting additional data never should be underestimated, there can be important costs in both time and money associated with additional data collection.

Manipulating X

The purpose of manipulating X is to change the values of X in a way that does not affect Yexceptthrough the changes induced in XÑin other words, in a way that has no direct effect on Y, only an indirect effect through the induced changes in X. If a mechanism can be found that accomplishes that purpose, then the changes in Y associated with the changes in X could be interpreted as the causal effect of X on Y.

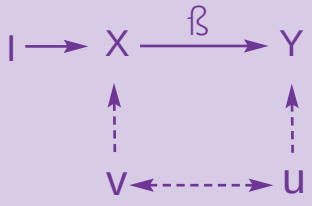
The most desirable mechanism for manipulating X independent of Y is random assignment of subjects to different values of X. A familiar example is assignment of subjects to a treatment (or ÒexperimentalÓ) group versus a control group. In that application, group membership is the variable X. When subjects are assigned randomly to the two groups (i.e., the two values of X) there is, in theory, no possibility that the assignment of group membership could be associated with any other variable that might affect the dependent variable (Y) independently of the effect through X.

In the absence of randomization, the analyst can look for another "manipulator." That manipulator might be one that the analyst controls, or one that occurs "naturally" in the data.

Of course, the theory can break down if subjects refuse to participate in the study or drop out during the course of the study for reasons that are related to the outcome variable.

In the absence of randomization, the analyst can look for another ÒmanipulatorÓ. That manipulator might be one that the analyst controls, or one that occurs ÒnaturallyÓ in the data. The general approach to Òmanipulation of XÓ is shown in Figure 9.

Figure 9:
Manipulating X

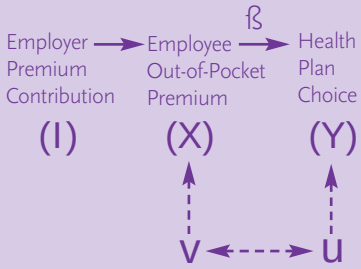


The goal is to identify a variable I that is correlated with X, but has no effect on Y other than through the changes it induces in X.

As an example, suppose that an analyst is trying to study the relationship between out-of-pocket premiums (X) and employee choice of health plan (Y) in firms that offer more than one plan. It is likely that in many settings, the overall premium for a health plan is affected by unobserved plan characteristics (e.g., quality-enhancing features) that also might affect the health planÕs appeal to employees. In that case, the estimated relationship between premiums and health plan choice would be biased, due to the effect of these unobserved plan characteristics that affect both premiums and choice.

Now suppose the model is altered so that employee choice of health plan is determined not by the total premium, but by the employeeÕs out-of-pocket premium, which equals the total premium minus any premium contribution the employer makes on ÒbehalfÓ of employees.[8] Suppose that the analyst is able to persuade the employer to change the employerÕs premium contribution[9] so that the employeesÕ out-of-pocket premium changes and suppose further, that it is plausible to assume that the change in the employerÕs premium contribution has no effect on health plan choice other than the effect through the employeesÕ out-of-pocket premiums. This model is shown in Figure 10. The employerÕs premium contribution is the manipulator or instrument, denoted I.

Figure 10:
Manipulating X: An Example



The employer's premium contribution affects the employee's out-of-pocket premium but has no other effect on the employee's choice of plan.

Assuming that changing the employerÕs premium contribution affects health plan choice only through changes in the employeesÕ out-of-pocket premiums, the analyst could compare the change in out-of-pocket premiums to the change in health plan enrollment and estimate the causal relationship between out-of-pocket premiums and health plan choice.

Ideally, the employerÕs premium contribution would be changed for a randomly chosen subset of sites or employees. Comparison of the treatment and control groups would allow the analyst to separate the effect of any ongoing time trend in health plan choice from the effect of the change in out-of-pocket premiums.

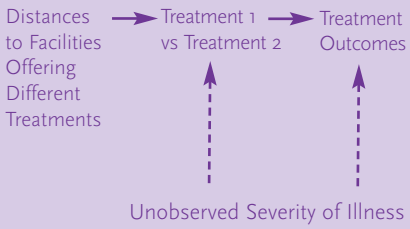
In the real world, variables that manipulate X with no direct effect on Y can be extremely difficult to identify. Saying that I affects X, but has no direct effect on Y is equivalent to saying that I affects X, but is not correlated with u. Because the requirements for I involve correlations with unobserved variables, it is understandable that the requirements are difficult to test.

This difficulty was illustrated by McClellan, McNeil, and Newhouse (1994) in a study that focused on the medical effectiveness of different treatments for acute myocardial infarction (AMI). McClellan, et al., realized that patients were not assigned randomly to different treatments, so there might be unobserved variables associated with the treatment received by the patient that also affected the outcome of treatment. The severity of the patientÕs medical condition was the variable most likely to be a source of omitted variables bias.

The authors were not in a position to manipulate directly any of the factors that affected assignment to treatment modalities, but they identified a variable that they argued produced different values of the treatment variable, but had no other effect on treatment outcome. The variable (I) was the difference in distances from the patientÕs home to facilities offering different types of treatments. The omitted variables of greatest concern to the authors were unobserved dimensions of severity of illness. Unobserved illness severity would be incorporated into the error term u in the Y equation. The authors demonstrated that distance was uncorrelated with observed measures of illness severity, and thus they concluded that it probably was uncorrelated with unobserved illness severity. The argument, then, was that Òdistance to facilityÓ was correlated with type of treatment (X), but not with u. In essence, the authors were arguing that Òdistance to facilityÓ assigned subjects randomly to different treatmentsÑrandom in the sense that the assignment process had no direct effect on treatment outcomes. McClellan, et al.Õs model is shown in Figure 11.

In the real world, variables that manipulate X with no direct effect on Y can be extremely difficult to identify.

Figure 11:
McClellan, et al.'s Model



When assignment to different values of X is not based on randomization, but on variation in a variable I, not all of the subjects may be at the same "risk" for being assigned to different values of X.

Having assigned subjects to values of X through a pseudo-random process, some method must be chosen to evaluate the effect of the treatment. There are two general approaches. The first is to compare the change in X to the change in Y for subjects with different values of the variable I. By comparing the differences in frequency of treatment types and differences in outcomes between patients who lived close to a facility offering a specific treatment and patients living far away from such facilities, the authors were able to calculate the causal effect of the treatment on outcomes. This approach is referred to as Òdifference in differences.Ó

When assignment to different values of X is not based on randomization, but on variation in a variable I, not all of the subjects may be at the same ÒriskÓ for being assigned to different values of X. This is an important feature of this estimation approach. When one relies on pseudo-randomization to assign subjects to different X values, the resulting estimates of the effect of X on Y apply to the ÒmarginalÓ patients (e.g., those at risk of being assigned different values of X). Harris and Remler (1998) discuss this issue in the context of health services research applications.

The second general estimation approach is a multivariate model that uses I variables and any additional predetermined variables in the model to predict X. Suppose we represent the equation for X as:

$$X = R + I + v$$

R is a vector of all the other predetermined variables in the model. and are coefficients and v is random error. Econometricians refer to this equation for X as a Òreduced formÓ equation. In this context, the predicted values of X are:

$$\text{Predicted } X = R^* + I^*$$

where the asterisk represents the estimated values of the coefficients.

The predicted value of X does not contain the error term v, because v has an expectation of zero for all values of X. Because the predicted value of X is correlated with the observed value of X, but has been purged of the troublesome source of correlation between v and u, it can be incorporated into an instrumental variable estimator that will yield consistent estimates of desired causal parameter, §. This approach is known as two-stage least squares (2SLS).[10]

The data requirement of the instrumental variable estimator is identical to the requirement for the change model or the difference-in-differences approach to estimation (i.e., a variable (I) that produces variation in X with no effect on Y, other than through the changes that I induces in X).

Modeling the correlation of u and v

The instrumental variable estimator based on the manipulation of X eliminates the correlation of u and v, and thus the correlation of X and u, by getting rid of v through differencing or through the use of predicted values of X in the instrumental variables estimator. Another class of estimation approaches deals with the problem of correlation of u and v by incorporating the correlation into the estimates of the causal parameter . When X is a discrete variable (e.g., membership in a treatment versus control group), these models are known in the econometrics literature as sample selection models.[11]

The possibility that unobserved variables affect both the choice of treatment versus control group, and also the subsequent outcome of interest, raises the possibility that v is correlated with u.

The essential problem in estimating the Y equation is that the expectation of the error term (u) is not equal to zero for the subjects with different values of X. For example, suppose the analyst is trying to determine the effect of membership in an HMO versus fee-for-service (FFS) health plan on utilization of services, but subjects self-selected into the two types of plans. The subjectÕs utilization data are observed only in the plan chosen by the subject. Suppose further that chronically ill individuals are more likely to join the FFS plan than the HMO, but chronic illness is unobserved to the analyst. In that case the expected value of utilization for HMO enrollees must be written as:

$$E(\text{UTILIZATION}^{FFS}) = X^{FFS} + E(u \mid_{\text{Choice of FFS}})$$

where $E(\text{UTILIZATION}^{FFS})$ means Òthe expected value of utilization for a subject enrolled in a FFS health plan.Ó $E(u \mid_{\text{Choice of FFS}})$ means Òthe expected value of the error term, given that the subject selected the FFS plan,Ó or in this example,

$$E(u \mid_{\text{Greater likelihood of chronic illness}}).$$

There are two common approaches to estimating the sample selection model. The first is a two-step approach. The first step is estimation of the X equation, that is, the equation that explains how subjects self-select into the groups (e.g., the treatment and control groups). From that first equation, a term is calculated that represents the expected value of the error term (u) given the sample selection rule. That term is added to the outcome equation of interest to correct for the fact that the error term, conditional on the self-selected sample, does not have a mean of zero. This approach is referred to as Òlimited information maximum likelihoodÓ or LIML.

The second estimation approach is simultaneous estimation of both the sample selection model and the outcome equation of interest using a maximum likelihood estimator. This approach is referred to as Òfull information maximum likelihoodÓ or FIML, and can be applied where X is either discrete or continuous.[12] The FIML approach requires an assumption about the joint distribution of u and v and has been criticized for that reason. Manning, et al. (1987) showed that the performance of the sample selection models depends crucially on the identification of at least

The possibility that unobserved variables affect both the choice of treatment versus control group, and also the subsequent outcome of interest, raises the possibility that v is correlated with u.

The third omitted variables approach, modeling the correlation of error terms across equations, also is not an option for correcting for reverse causality bias. The bias from reverse causality arises whether or not errors are correlated across equations.

one variable that affects sample selection, but otherwise is unassociated with the outcome variable in the equation of interest. That, of course, is exactly the same data requirement as for the difference in differences, or instrumental variables approach.

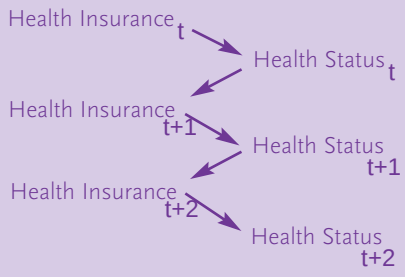
Reverse causality

Even though the source of the bias caused by reverse causality is different from that of omitted variables, Figure 8 shows that the result is the same: the explanatory variable X is correlated with the error term (u) in the Y equation. Thus, it is not surprising that the approaches to the two problems share some common ground. However, not all the approaches to omitted variables bias are helpful in addressing the problem of reverse causality. Collection of additional data will not solve the problem of reverse causality, for example, even if the additional data are defined over more finely differentiated units of time that allow the analyst to convert the model from the problem shown in Figure 8 to the type shown in Figure 12.

In Figure 12, data on health insurance (X) and health status (Y) are taken from successive time periods so that the question of causality for each variable in each time period is settled by the antecedence of the other variable. The structure of the model in Figure 12 implies that it is impossible for health status during time period t (e.g., a contract year for health insurance) to affect health insurance status during the current time period t. Health status at time t affects health insurance, but only in the next time period t+1 (e.g., the following contract year). Similarly, the diagram implies that health insurance at time t can affect health status during time t, perhaps through improved access to care. But if all this information truly is known to the analyst, the β_{ss} coefficient could have been declared to be θ in Figure 7, assuming that health insurance and health status were measured in the same time period in that Figure.

The third omitted variables approach, modeling the correlation of error terms across equations, also is not an option for correcting for reverse causality bias. The bias from reverse causality arises whether or not errors are correlated across equations (as shown in Figure 8).

Figure 12:
The Effect of Additional Time Periods on the Problem of Reverse Causality



Conclusions continued

Econometricians are perhaps the most optimistic of causal modelers, and even among econometricians, there are those who have grave misgivings about the whole causal modeling enterprise. This paper has provided an introduction to only two of many specific threats to causal modelingÑomitted variables bias and reverse causality. The conclusion probably should re-emphasize the difficulties associated with analytic approaches to those problems.

First, it often is impossible, as well as unethical, to assign subjects randomly to different values of the X variable. Thus, the most effective method of manipulating the values of X often is removed from our choice set. Even when randomization is possible, however, subjects often refuse to participate or drop out of the study before the observation period is complete, or the control group becomes contaminated in some way.

In the absence of randomization, the identification of mechanisms to manipulate X that meet the prerequisite conditions (correlation with X, but not with u) can be extremely challenging, and often impossible. Even when a reliable mechanism (I) is identified, the choice of estimation methods can be controversial, as well. For the past several years, a well-known health economics research journal has refused to publish any article using sample selection models, despite the fact that the 2000 Nobel Prize in economics was shared by one of the principal developers of the method. Perhaps the best advice that can be given at this point to those new to the field is to:

a. Read as extensively as possible the literature on causal modeling and causal inference.
b. Try to read from more than one discipline.
c. Try to understand both the historical perspective of different disciplines, as well as new developments.
d. Be prepared to encounter a variety of attitudes toward causal modeling.
e. Be as clear as possible about your own intentions, models, and results.

Endnotes

1 Throughout the paper, cross-sectional data means data that are taken from the same time period, so that only one time period is observed for each subject. Econometricians refer to data on multiple subjects over multiple time periods as panel data.

2 In this case, ÒlinearÓ means linear in both the variables and the parameters (i.e.,).

3 Throughout the paper, Ònon-experimentalÓ is synonymous with Ònon-randomized,Ó and refers to data taken from settings in which subjects were not assigned randomly to different values of X.

4 Throughout the paper, we omit the constant or ÒinterceptÓ term in the regression equations for simplicity.

5 If errors were correlated across equations, butÒtype of health planÓ did not affect Òhealth expenditures,Ó there would be no bias in the estimated coefficients $_z$ or $_x$, but the standard errors of the coefficients estimated from ordinary least squares would be biased. Econometricians call that model Òseemingly unrelated regressionsÓ because the only connection between the equations for Òtype of health planÓ and Òhealth expendituresÓ is the correlation of the error terms.

6 Throughout the paper, Òbiased coefficientsÓ refers to biased coefficients estimated from ordinary least squares regressions.

7 ÒNaturallyÓ occurring I variables that assign subjects to values of X in a pseudo-random way (i.e., one that has no direct effect on the outcome of interest) has given rise to the term Ònatural experimentÓ in the econometrics literature. A change in the tax laws that affect the after-tax price of a good or service in one time period versus another or for one group versus another would be an example of a natural experiment.

8 Economic theory predicts that in the long run, employees pay the full cost of health insurance premiums and other fringe benefits out of foregone wages.

9 One reason the analyst might be successful in persuading the employer to change the firmÕs premium contribution is that the analyst is the employer.

10 The instrumental variables estimator does not simply substitute the predicted values of X for X. In matrix notation, the estimator is $(X'Z)^{-1}(X'Y)$, rather than $(X'X)^{-1}(X'Y)$ where Z is the vector of explanatory variables with the predicted value of X substituted for the actual value of X. The estimated standard errors of the coefficients also are corrected for the presence of predicted, rather than actual, values of X among the explanatory variables.

11 When X is a continuous variable, they are known as full-information maximum likelihood or FIML models, but FIML also is used to refer to a specific estimation approach for a discrete X.

12 When X is a continuous variable and the error terms in each equation are normally distributed, the three stage least squares (3SLS) estimator is equivalent to the FIML estimator (Greene, 2000, p. 695).

References

Agency for Health Care Policy and Research (ACHPR). Proceedings from a conference: Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data, Tucson, Arizona, April 8-10, 1987. Sechrest, L., B. Starfield and J. Bunker, eds. U.S. Department of Health and Human Services Publication Number (PHS) 90-3454 (May 1990).

Christianson, J., A. Riedel, D. Abelson, R. Hamer, D. Knutson and R. Taylor. Managed Care and Treatment of Chronic Illness. Sage Publications: Thousand Oaks, CA (2001).

Greene, W. Econometric Analysis. Prentice-Hall, Inc.: Upper Saddle River, NJ (2000).

Heckman, J. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," National Bureau of Economic Research Working Paper 7333, available at www.nber.org/papers/w7333. Cambridge, MA (September 1999).

Harris, K. and D. Remler. "Who is the Marginal Patient? Understanding Instrumental Variables Estimates of Treatment Effects," Health Services Research 33:5 (December 1998, Part 1) pp. 1337-1360.

Heise, D. Causal Analysis. Wiley and Sons: New York (1975).

Kennedy, P. Guide to Econometrics. 4th Edition. MIT Press: Cambridge, MA (1998).

Maddala, G.S. Introduction to Econometrics. Macmillan Publishing Company: New York (1988).

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

McClellan, M., B. McNeil, and J. Newhouse. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" Journal of the American Medical Association 272:11 (September 21, 1994) pp. 859-866.

Pearl, J. Causality: Models, Reasoning and Inference. Cambridge University Press: (March 2000).

Salmon, W. Causality and Explanation. Oxford University Press: New York (1998).

Webster's New Collegiate Dictionary. G. C. Merriam Company: Springfield, MA (1977).

About the Authors

Bryan Dowd is professor and director of graduate studies in the Division of Health Services Research and Policy at the University of Minnesota where he teaches health services research methods in the doctoral and masterÕs programs. His primary research interests are markets for health insurance and health care services and evaluation of non-experimental data. Recently, he has directed studies of health plan enrollment and disenrollment, health outcomes, and employersÕ health insurance purchasing strategies.

Robert Town is an assistant professor in the Division of Health Services Research and Policy at the University of Minnesota where he teaches health economics in the doctoral program. His primary research focus is on competition in the health care marketplace. Dr. Town has studied the impact of network formation in hospitals competition, the role of competition in determining hospital quality, and the appropriate antitrust policy in health care and health insurance markets.

About the Program

This report was prepared by AcademyHealth under The Robert Wood Johnson FoundationÕs Changes in Health Care Financing and Organization (HCFO) program. HCFO encourages the development of policy analysis, research, evaluation, and demonstration projects that provide policy leaders with timely information on health care policy, financing, and market developments.

Anne K. Gauthier, Program Director

Deborah L. Rogal, Deputy Director


AcademyHealth
Advancing Research, Policy and Practice