An economist's primer to graphical models and causal identification

J. Wesley Burnett[*] and Calvin Blackwell[†]

June 1, 2021

Abstract

This paper offers an accessible discussion of graphical causal models and how such models can be used to identify causal relations. A graphical causal model is a visual representation of a researcher's qualitative assumptions. There is growing interest among economic analysts in using this framework to analyze cause-and-effect relationships. To explain how DAGs work, we offer numerous illustrations to help the reader disentangle the complex pathways that link the treatment to the outcome. In addition to graphical illustrations, we include an intuitive explanation of the sufficient conditions needed to ensure that the analyst's story conforms to his or her claims of causation. To demonstrate how DAGs can be used in practice, we develop a Monte Carlo simulation to estimate an average treatment effect using an identification strategy called the front-door criterion.

Keywords: Causal analysis; Graphical causal models; Directed acyclic graphs; Front-door criterion; Backdoor criterion.

JEL Codes: B41; C10; C52.

[*] Research Economist, U.S. Department of Agriculture, Economic Research Service, 805 Pennsylvania Avenue, Kansas City, MO 64105, wesley.burnett@usda.gov.

[†] Professor, Department of Economics, College of Charleston, 5 Liberty Street, 403 Beatty Center, Charleston, SC 29401, blackwellc@cofc.edu.

1 Introduction

This paper offers a contemporary introduction and overview of graphical analysis and how this approach can be used to help validate causal relations within non-experimental or observational data. A graphical causal model is a visual depiction of the assumptions underlying a researcher's model of cause-and-effect. This approach is attractive because it is compatible with potential outcomes, which has become one of the most prominent methodological approaches within the social sciences. Potential outcomes (or, the "Neyman-Rubin-Holland model") is a statistical analysis, based on observational data, of causal effects (Neyman, 1923 [1990]; Rubin, 1974; Holland, 1986). For the sake of space, we provide only a brief discussion of the potential outcomes framework in an online appendix.[1] Complementary to potential outcomes, a graphical causal model provides simple rules that map causal assumptions to associations and independencies within the underlying data.

Against this backdrop, we outline the purposes for using a graphical causal model. One, causal graphs help to determine the identifiability of causal effects from observational data. Two, graphical models can be used to clearly depict and outline the assumptions within empirical modeling. Historically, analysts have explained their modeling assumptions through often long narratives and mathematical representations. As an alternative, graphical models offer all of the assumptions in one snapshot, which can be far easier for a reader to understand. This assertion is based on decades of research within the psychology literature, which shows that pictorial illustrations help to improve a person's learning from text (Carney & Levin,

[1] This paper is written such that it is assumed the reader has little or no prior knowledge of potential outcomes. Nonetheless, a basic understanding of the framework may enrich the reader's understanding of graphical causal models.

2002). Put simply, this approach helps to determine which effects warrant causal interpretation.

This paper explores graphical analysis in the context of directed acyclic graphs (DAGs).[2] Similar to the other social science disciplines, economic researchers are demonstrating a growing interest in using causal graphs for empirical analysis. This is perhaps due to the fact that graphs in causal modeling were only recently revitalized by Pearl (1988, 1993, 1995). Pertinent discussions of causal graphs within the economic literature include Spirtes (1993, 2005), Heckman & Pinto (2015), Cunningham (2021), Barenboim & Hünermund (2019), Imbens (2020), and Bellemare *et al.* (2020). Imbens (2020) contends that DAGs deserve the attention of all researchers and particularly for those appealing to causal inference.[3]

We contribute to the literature by emphasizing two subtle but incredibly important takeaways from graphical causal modeling. First, causal graphs are helpful for designing a credible identification strategy. However, causal *inference* is not the same as *identification*, which is an exploration of what we can learn about modeling parameters given an infinite amount of data(Barenboim & Hünermund, 2019; Manski, 2007). In constrast, inference pertains to what we do not observe (parameters) from what we do observe (data). Thus, identification is a necessary but not sufficient condition for causal analysis. In addition to identifying the causal parameters, economic theory and prior knowledge are needed for the analyst to draw causal inference.

Second, graphical causal modeling helps to reduce many forms of estimation

[2] Our outline of DAGs should not to be construed as an exhaustive treatment. For a comprehensive review of DAGs, the reader is referred to Morgan & Winship (2015), Pearl (2009), (Pearl, 2018), and Imbens (2020), among others.

[3] Heckman & Pinto (2015) seem amenable to DAG analyses, but criticize the approach as not being able to accommodate nonrecursive simultaneous equations models.

bias, including over-control bias. To wit, an analyst needs to decide whether or not the addition of a control variable to a regression equation will help to yield an unbiased estimate for the causal parameter of interest. The challenge is to determine which variables need to be included so that omitted variable bias can be minimized or eliminated. The conventional wisdom within applied econometrics is to include as many potential confounders as possible to control for omitted variable bias. Cinelli *et al.* (2021) define the inclusion of confounders, in a regression equation, as "good controls." However, if the researcher is not careful, she may unintentionally include deconfounders or "bad controls." (Angrist & Pischke, 2009; Cinelli *et al.* , 2021). The inclusion of bad controls is a form of over-control bias, which can exacerbate the estimation bias for the causal parameter of interest. Issues of omitted variable bias are well known among applied economists, but comparatively less attention has been given to over-control bias (Elwert & Winship, 2014; Grätz, 2019). Graphical modeling can aid in reducing both forms of bias and help ensure a parsimonious specification of the data generating process.

Differing from Imbens (2020) and Heckman & Pinto (2015), our paper is more of a practitioner's guide to graphical causal modeling. Motivated by the work of Pearl *et al.* (2016), we seek to intuitively explain the main concepts and minimize the discussion of theory. Our discussion is tailored more toward applied economic research using observational data, and we offer a more in-depth discussion of the graphical identification criteria (front door and backdoor criterion). Lastly, we offer an introduction to a user-friendly, open-source program (*dagitty.net*), which automates the criteria formulation and determines if causal identification is possible given an analyst's proposed model.

2 Background

Directed acyclic graphs have roots in structural equation models, social science path models, and Bayesian networks. Pearl (1995, 2000, 2009) synthesized and generalized these approaches to develop DAGs for causal inference. We will follow Pearl (1995, 2009) and interpret DAGs as nonparametric structural equation models.

Our goal is to examine DAGs based on fully parametric regression models within the context of applied economics. DAGs can be generalized to different types of regression models including cross-sectional, panel data, time series, and count models, among others. Causal graphical models encode the analyst's qualitative causal assumptions about the data generating process in the population (Elwert, 2013). Additionally, the approach helps to derive implications of a causal model such as conditional independence assumptions.[4] Perhaps the most important purpose of DAGs is to aid in causal identification. Identification is the possibility of separating causal from non-causal associations with ideal data.

2.1 Terminology and Preliminaries

Direced acyclic graphs consist of only three elements: variables, arrows, and missing arrows (Elwert, 2013). Variables – sometimes referred to as nodes, or vertices – may be observed or unobserved. Arrows – or directed edges or arcs – represent possible causal effects and the directionality of the effects. It is important to note that the arrows order the variables in time – that is, it is assumed that the cause has to precede the effect in time. For example, the DAG in Figure 1 implies that the variable X precedes the occurrence of the variable I, which in turn, precedes the

[4] The term "conditioning" is defined as introducing information about a variable into an analysis; e.g., through sample selection, stratification, or regression control (Elwert & Winship, 2014).

occurrence of variable S. In notation, the order of occurrence can be written as: $X \to I \to S$. Finally, missing arrows represent sharp assumptions about absent causal effects. Additional terminology for DAG analysis, including a description of the ancestral structure, is provided in the online appendix.

[Figure 1 will go about here.]

To fix ideas, Figure 1 offers a DAG representing a relatively simple data generating process. First, it is worth noting that the DAG does not contain independent error terms for each of the variables. The error terms (or, idiosyncratic direct causes) are suppressed because the error terms generally do not aid in identification (Györfi *et al.* , 2002). For illustrative purposes, the DAG in Figure 1 represents an example of returns to schooling, which examines the causal effect of education on a person's later earnings (Mincer, 1958, 1974; Angrist & Krueger, 1991; Staiger & Stock, 1997; Goldin & Katz, 2000; Oreopoulos & Salvanes, 2011; Stephens Jr & Yang, 2014; Feigenbaum & Tan, 2019; Clay *et al.* , 2021). We use returns-to-schooling as a running example throughout the remainder of the manuscript.

For Figure 1, we assume that X denotes ancestral heritage (or social origin) and characteristics (race, ethnicity, place of origin, etc.), I presents parents' income, P denotes parents' education, S (the treatment variable) represents years of schooling, Y (the outcome variable) is current earnings, and U denotes an unobserved confounder. The primary interest in this case would be to identify the causal effect of years of schooling on current earnings. To aid in interpretation, we illustrate observed variables with solid nodes (●), whereas unobserved variables are represented by empty nodes (○).

One of the primary strengths of the causal graphical approach, is its ability to offer an intuitive narrative in a single snapshot. For example, Figure 1 offers a depiction of the researcher's assumptions about how schooling affects earnings. Within this figure, there are multiple possible causal pathways. One potential causal chain is represented in notation as: $X \rightarrow I \rightarrow S \rightarrow Y$. In words, the chain indicates that ancestral heritage affects parents' income ($X \rightarrow I$), and, in turn, parents' income affects years of schooling ($I \rightarrow S$). Finally, years of schooling attained causes current earnings ($S \rightarrow Y$).

In addition to directed causal paths, such as $X \rightarrow I \rightarrow S \rightarrow Y$, the DAG also includes potential backdoor pathways that would affect the researcher's ability to identify causation. A backdoor path is any non-causal pathway, that unless controlled for, may lead to forms of spurious association. For example, the model assumes that the unobservable confounder affects both ancestral heritage and one's current earnings. In notation, this backdoor path is: $X \leftarrow U \rightarrow Y$. The unobserved confounder, U, is explicitly included in the model as a potential causal factor, as the omission of this variable would yield, at minimum, a biased treatment effect estimate. An analyst would presumably identify the confounding factor by bring prior, substantive knowledge (such as economic theory) to the problem.

A researcher who is unfamiliar with DAGs is likely to construct a mental narrative as she is conceptualizing a problem, such as returns-to-schooling. When presenting the results, the researcher often has to explain in words her qualitative assumptions for causality, and depending on the complexity of the model, this discussion may be quite lengthy. A DAG is helpful in this context as it clarifies the researcher's thought process (mental constructs of the problem) and the final product provides an pictorial road map of the researcher's causal claims. As an added bonus, this map could aid readers to more quickly and easily understand the researcher's assumptions.

Another strength of DAG analysis is that the models are fully nonparametric constructs (Pearl, 1995, 2009). In other words, the distribution of variables (or nodes) in the DAG can be continuous or discrete, among others. Further, the functional form of the direct causal effects (the arrows or arcs) can be linear or nonlinear, among others. Since DAGs are nonparametric, the functional forms can also accommodate interaction terms or higher order polynomials of the variables (without any explicit representation of such within the DAG). In general, DAGs do not place restrictions on the function form or parametric specifications of models (Heckman & Pinto, 2015). DAGs also allow for effects heterogeneity; that is, the effects can freely vary across observational units. Lastly, this approach does not constrain the magnitude or sign on an estimated causal effect (Elwert, 2013).

2.2 Causal and Non-Causal Pathways

The arcs within the DAG are generally referred to as causal and non-causal "paths" (Elwert, 2013). More specifically, a path is a sequence of non-intersecting adjacent edges. Non-intersecting means that a singular path cannot cross a node more than once. DAGs are *acyclic* in that they contain no directed cycles; that is, one cannot trace a sequence of arcs and arrive where one started. A path in which all arrows point away from the treatment (years of schooling, S) to the outcome (current earnings, Y) is defined as a "causal path." The total causal path of a treatment on an outcome includes all of the causal paths connecting the two. A "non-causal path" consists of any path in which at least one directed arc points against the flow of time (e.g., $X \leftarrow U \rightarrow Y$). Non-causal paths represent potential sources of spurious association between treatment and outcome. Thus, DAGs can be thought of as statistical models.

A common *misconception* within the graphical causal modeling approach is that providing more directed arcs will lead to more robust inference. The directed arcs represent possible causal pathways. However, if one includes a directed arc from each node to every other node within the DAG, then the model would be rendered nonsensical; that is, it would be challenging to draw inference from the model. Elwert (2013) argues that adding arrows to a given node (i.e., relaxing assumptions) *never* helps nonparametric identification, but adding nodes (variables) may help improve parametric identification.

Consider Figure 2, which illustrates a complete DAG – a graph with arcs going to every variable within the model. Although this model contains fewer assumptions than Figure 1, it would be nearly impossible to draw any causal inference from such a model.

[Figure 2 will go about here.]

Hence, the *truer* assumptions in the model are determined by the *missing arcs* within the DAG (Elwert, 2013). These missing arcs represent assumptions of no causal effect. Thus, the missing arcs encode the causal pathways within the model, whereas the directed arcs represent ignorance (Morgan & Winship, 2015). Put differently, the missing arcs represent the *exclusion restrictions* within the model. For example, the missing arc between X and P (in Figure 1), which represents the analyst's assumption that ancestral heritage does not directly affect parents' education. Thus, missing arcs create contestable hypotheses about the relationships among the variables.

2.3 DAGs and Instrumental Variables

Perhaps the strongest assumption in Figure 1, especially for the reader who is familiar with this literature, is a lack of directional arrow between the unobserved confounder, U, and the years of schooling, S. As posited by Angrist & Krueger (1991), a person's latent 'abilities' (or motivations) may have a considerable impact on years of schooling attained and current earnings.

[Figure 3 will go about here.]

Figure 3 offers a DAG that includes ability as an unobserved confounder that affects years of school attained. In other words, there is now a directed arc going from U to S. Adding this simple, but consequential assumption, introduces a common confounder on returns to schooling and current earnings (i.e., $S \leftarrow U \rightarrow Y$) that renders the causal effect unidentifiable (at least through adjustment).

However, an exogenous variable Z was added to the DAG (Figure 3), in which it is assumed that Z only affects years of schooling attained. As is well documented within the economics literature, Z can be used as instrument for S. If this instrument is valid, then the researcher can appeal to causal inference by using an instrumental variables (IV) estimation approach. As an example, Angrist & Krueger (1991) used quarter-of-birth and mandatory schooling laws as an instrument for years of schooling.

If the researcher is to appeal to causal inference within an IV approach, then the DAG representation now implies the following sufficient conditions: (i) X is independent of P; (ii) X is independent of Z; (iii) I is independent of Z; and, (iv) P is independent of Z. The online appendix offers a brief explanation of the (conditional) independence assumptions (i.e., the ignorability or conditional ignorability assumptions) required for causal inference. In short, these assumptions require that the potential outcomes are “statistically independent" (or, paraphrased as independent) of the treatment in order to identify a causal relationship.

The first sufficient condition implies that ancestral heritage is independent of parents’ education. This first condition is quite strict. For example, consider a researcher who has data on grandparents’ education. If parents’ education is assumed to have an affect on the observed unit’s years of schooling ($P \rightarrow S$ in Figure 3), then it is likely that grandparents’ education would have an affect on parents’ education ($X \rightarrow P$). Yet, notice that there is no directed arc from X to P in Figure 3. Thus, the first condition may be violated, and the burden of proof is on the analyst to convince the reader otherwise.

The second through fourth conditions (the exclusion restrictions) imply that the instrumental variable is independent of ancestral heritage, parents’ income, and parents’ education. The last three conditions are required for the instrument to be exogenous. If any of these conditions are potentially violated, then the analyst may not be able to identify the causal effect of schooling on earning. In which case, the model should arguably be revised.

Beyond the DAG literature, but well documented within the economics literature, the IV approach may be problematic if the instruments are weak (Bound *et al.* , 1995). That is, if the instrument, Z, explains little of the variation in the endogenous variable, S. This is a prime example in which graphical causal analysis can help an analyst to identify an effect (from instrumented years of schooling to earnings), but the analyst must appeal to prior knowledge (or economic theory) to infer that the relationship is causal.

2.4 Causation, Confounding, and Endogenous Selection

Now that we have the basic terminology, we will turn our attention to the types association found within the graphical causal modeling framework. There are three types of structures from which all DAGs can be constructed: chains, forks, and inverted forks (Morgan & Winship, 2015; Pearl, 2009; Elwert, 2013). Consider the hypothetical causal chain represented in Figure 4a. In this example, the researcher only has data on years of schooling attained, S, a new variable of participation in a job training program, T, and current wages, Y. In notation, this path is $S \to T \to Y$. Or, in words, years of schooling attained affect performance within the training program, which, in turn, causes current earnings.

A fork is represented in Figure 4b, and can be written as: $I \leftarrow P \to S$. Put differently, parents' education affects both parents' income and the observational unit's years of schooling attained.

An inverted fork is represented in Figure 4c, and can written as: $X \to I \leftarrow P$. The inverted fork implies that parents' education and ancestral heritage separately affect parents' income.

[Figure 4 will go about here.]

The structures in Figure 4 directly correspond to causation, confounding, and endogenous selection (Elwert, 2013). The chain of causation, in Figure 4a, implies that S is not a direct cause of Y, but rather S now indirectly causes Y through the mediating factor of the training program, T . In terms of notation, the first clause would be written as $Y \not\perp\!\!\!\perp S$, and the second clause would be written as $Y \perp\!\!\!\perp S \mid T$. The term "$\perp\!\!\!\perp$" implies statistical independence, whereas the term "$\not\perp\!\!\!\perp$" implies that two variable are *not* statistically independent of one another.

Common cause confounding, as represented in the context of Figure 4b, implies that I (parents' income) does not directly cause S (years of schooling).[5] Yet, I may appear to be associated with S as the two have the common cause of P (parents' education), or in notation form: $I \not\perp\!\!\!\perp S$.[6] However, I and S are statistically independent after conditioning on P: $S \perp\!\!\!\perp I \mid P$. Put different, if we were to conduct an auxiliary regression of years of schooling (dependent variable) on parents' income (independent variable), then we would likely find a highly statistically significant relationship between the two, but this finding would be spurious or non-causal. On the other hand, if we ran the same regression but also added parents' education as an explanatory variable (in addition to parents' income), then the new regression should now yield a non-spurious estimated effect of parents' income on years of schooling.

The inverted fork, Figure 4c, implies that P (parents' education) is marginally independent of X (ancestral heritage) or in notation: $P \perp\!\!\!\perp X$. But, P is *not* independent of X if the researcher conditions on I; or, in notation: $P \not\perp\!\!\!\perp X \mid I$.

2.5 Over-control bias, Confounding bias, and Endogenous selection

Elwert (2013) defines three types of bias associated with conditioning as: "over-control bias," "confounding bias," and "endogenous selection." Over-control bias is represented in Figure 6a, where the square box, □, denotes conditioning on the variable S. Consider a case in which the researcher is interested in ascertaining

[5] This particular implication does not apply in Figure 1 as parents' income, I, is assumed to directly cause years of schooling attained, S.

[6] An apt example of this type of confounding is the relationship between ice cream sales and an increase in an area's crime rate. The common confounder is warmer temperatures within the area. In the absence of controlling for temperature, it may appear that ice cream sales and crime are related to one another.

how parents' income, I affects a person's current earnings, Y. The causal chain in Figure 6a implies that I is marginally associated with S – two variables are marginally associated if one directly or indirectly causes the other. As a result, if the researcher conditions on S, then it would block, or control away, the association between I and Y (Elwert, 2013). In other words, the conditional between I and Y given S would potentially *not* identify the effect of parents' income on current earnings. In terms of a regression of Y on I and S, over control bias implies the years-of-schooling variable would potentially absorb all of the variation from the parents' income variable. Therefore, the researcher would not necessarily be able to identify an unbiased affect of parents' income on current earnings, given this particular specification.

[Figure 6 will go about here.]

Confounding bias occurs when two variables share a common cause. In the context of Figure 4b, the true underlying association between I (parents' income) and S (years of schooling) is only identified after one conditions on P (parents' education). In the absence of conditioning on P, any association between I and S is spurious (as mentioned above) because the association does not identify a causal effect. On the other hand, the conditional association between I and S given P would identify the causal affect of I on S, but in the explicit representation within Figure 4b, the causal effect in this case would equal zero.

Conditioning on a collider or its descendant (next variable in the causal chain that is marginally associated with the collider) creates *endogenous selection*, which is another form of spurious association (Elwert, 2013). A collider is a node with two arrows pointing to it, as in Figure 4c. A collider is always a non-causal path.

Put differently, the spurious association occurs between two variables (that would otherwise be marginally independent) after conditioning on the collider. Another way to see the spurious correlation is through Figure 6b, where again the square box around I represents conditioning on that variable, and the dashed line illustrates the spurious correlation between P and X introduced by conditioning on the collider I. Conditioning on the descendent of a collider, as in Figure 6c, also induces spurious association.

The next subsection offers an intuitive example of endogenous selection (or over-control bias specifically) within the social science literature.

2.5.1 Example of over-control bias: Intergenerational Mobility

To illustrate over-control bias, we will briefly explore an alternative research question given the same data and DAG (Figure 1) for the returns-to-schooling example. As opposed to focusing on the returns of schooling, we instead will examine intergenerational mobility – changes in status between individuals of different generations of family (Grätz, 2019; Corak, 2013). In the case of intergenerational mobility, a parent's income, I, becomes the treatment variable of interest, and a child's income, Y, remains the outcome of interest. Since this research question is slightly different, we offer the alternative DAG in Figure 5.

The reader will notice that the DAG, in Figure 5, is almost identical to the previous ones. Only now, there are two unobserved confounders, U_1 and U_2. The sociology literature uses a slightly different terminology by referring to X as "social origin" (Grätz, 2019); otherwise, the interpretation is nearly identical to ancestral heritage (our running definition of X thus far). The inclusion of the second unobserved confounder is not a new addition to the model, only in this case

U_2 can be interpreted explicitly as latent ability. Prior to this example, we allowed for unobserved ability to be subsumed into the single unobserved confounder (U in Figure 1).

The only difference with the intergenerational mobility example is that now analyst is interested in identifying the causal effect of parent's income on child's income. The data generating process remains the same as in our previous running example, but this slight change in focus renders an unidentifiable cause and effect. In this case, controlling for child's educational attainment introduces over-control bias (Grätz, 2019) – hence, the box around schooling, S, to illustrate that the analyst is conditioning on it.

In other words, a child's unobserved ability (U_2) is assumed to affect both attained schooling and income. Schooling is then a collider on the causal path from parent's income to child's income, and if one controls for schooling then the causal affect cannot be identified because doing so opens a non-causal path. This is referred to as collider bias or over-control bias. Alternatively, if the researcher does *not* control for schooling (i.e., do not include schooling as a covariate in the context of regression analysis), then the causal effect *can be identified* after controlling for parent's education, and social origin. The only contestable hypothesis (which is arguably problematic) is that social origin, X, is statistically independent of parent's education, P.

[Figure 5 will go about here.]

3 Sufficient Conditions for Causal Identification within Graphical Models

The general lesson from the last section is that all associations between variables in a DAG are transmitted along paths, yet not all paths transmit association (Elwert, 2013). Whether the path transmits association depends on the direction of the arrows and what variables the analyst conditions on. As these associations can sometimes create confusion, Pearl (1988) developed a set of rules or criteria for causal identification within graphical causal models.

3.1 Backdoor Criterion

These criteria consist of the “front-door criterion" and the “backdoor criterion" (Pearl, 2009). The backdoor criterion (BD) is important because the researcher does not want to open any backdoor paths so as to create any non-causal correlations between the causal variable of interest and the outcome (Cunningham, 2021). The BD is defined as follows, which is paraphrased for illustrative purposes.

Definition 1 *(Pearl* et al.*, 2016, p. 61) Backdoor criterion:*
Given a treatment variable S and an outcome variable Y, the researcher conditions on all other variables such that:

1. *All spurious paths are blocked between S and Y;*
2. *All direct paths from S to Y are left unperturbed; and,*
3. *No new spurious paths are created.*

As an example of the BC, consider the simplified version of our returns-to-schooling story offered in Figure 7a. In this instance, the researcher is still primarily interested in identifying the effect of years of schooling, S, on current earnings, Y. Yet, the researcher only has additional information on parents' income, I, and the researcher knows there is an unobserved confounder, U, that affects both years of schooling and parents' income. Given this specific representation, the BD would be satisfied as parents' income is not a descendant of years of schooling, and parents' income blocks the backdoor path created by the unobserved confounder: $S \leftarrow U \rightarrow I \rightarrow Y$. So if the causal story conforms to the graph, then adjusting (or conditioning) for I is sufficient to yield the causal effect of S on Y.

[Figure 7 will go about here.]

Shpitser & VanderWeele (2011) define the "adjustment criterion," which is a generalization of Pearl's (1993) BD. The adjustment criterion is a very practical sufficient condition for causal identification, which is defined as: (i) close all non-causal paths; and, (ii) leave open all causal paths.

3.2 Front-door Criterion

The backdoor criterion provides a simple method for identifying sets of covariates that should be adjusted in order to appeal to causal inference; but, the BD does not exhaust all ways of estimating causal effects. To wit, the front-door criterion (FD) potentially allows for alternative forms of adjustment in order to appeal to causal identification.

Consider the simplified version of our returns-to-schooling example offered in Figure 7b. In this case, the researcher only has data on the years of schooling attained, S, and current earnings, Y. Moreover, the researcher brings substantive knowledge to the problem and identifies an unobserved confounder, U, that affects both years of schooling and earnings. With this particular specification, the causal effect of schooling on earnings cannot be identified as the confounder is unobserved, and there is nothing to block the backdoor path: $S \leftarrow U \rightarrow Y$. Put another way, the backdoor criterion fails, and any estimated relationship between S and Y is spurious.

On the other hand, consider the slightly augmented case in Figure 7c, in which this researcher also has individual level data on performance within a training program, T. In this case, the training program serves as a mechanism or mediator along the causal path of years of schooling to current earnings: $S \rightarrow T \rightarrow Y$. And, as before, the researcher is still interested in identifying the causal effect of schooling on earnings. The unobserved confounder, U, still affects years of schooling, S and current earnings, Y.

With this particular specification, the front-door criterion is defined as follows (Pearl, 1995, 2000; Bellemare *et al.* , 2020).

Definition 2 *(Pearl, 1995, 2000) Front-door criterion:*
Given a treatment variable S, an outcome variable Y, and a mediator T, the criterion is satisfied given the following conditions:

1. *T intersects all causal paths from S to Y;*
2. *T is the only mediator or mechanism along the path (i.e., there are no backdoor paths); and,*
3. *Every backdoor path between T and Y is blocked by S.*

Once the conditions for the FD are satisfied, the diagram can be decomposed into two causally identified relationships.[7] The first causal relationship $S \rightarrow T$ is

[7] One can also interpret this as applying the BD to the two separate causal relationships.

identified because the unobserved confounder affects years of schooling but not participation in the training program. The second causal relationship $T \to Y$ is identified because the unobserved confounder affects current earnings but, again, does not affect participation in the training program.

Given the DAG in Figure 7c and satisfaction of the FD, the estimation procedure is quite intuitive, in that it consists of two reduced-form regression equations. This approach corresponds to the “product method," due to estimating the indirect effect as the product of two estimated coefficients from each equation (Baron & Kenny, 1986). The first is a regression of T on S: $E[T \mid S = s] = \beta_0 + \beta_1 s$. The second is a regression of Y on T: $E[Y \mid S = s, T = t] = \theta_0 + \theta_1 t + \theta_2 s$. Based on these two equations, and satisfaction of the FD, this approach yields a key parameter estimate, $\hat{\theta_1}$ and $\hat{\beta_1}$, from each equation. (The caret symbol “$\wedge$" over the parameters is notation indicating that the parameters are estimated.) The average treatment effect is then derived from the product of these two parameter estimates: $\widehat{\beta_1 \cdot \theta_1}$ (Pearl, 1993). The causal interpretation of the average treatment effect in this example assumes *no* unobserved mediator-outcome confounding; i.e., $T \perp\!\!\!\perp Y \mid S$.

3.3 Graphical Identification Criteria

The front-door and backdoor criterion are collectively referred to as the “graphical identification criteria" (Pearl, 1988, 2009). The graphical identification criteria provide the sufficient conditions (or contestable implications) based on the required statistical independencies for causal identification.

The contestable implications within the DAG are merely a reflection of the criterion needed to nonparametrically identify the causal effect. Specifically, these implications are a set of hypotheses of conditional independence constraints implied by the model. They are contestable in that implications can be refuted as invalid. Testing these implications is challenging in that they are often based on potential outcomes that are not always observed. Nevertheless, if any of the testable implications can be refuted, then the specified DAG arguably violates identification of the causal relation.

As an example, let's revisit the DAG representation of returns-to-schooling offered in Figure 1. Again, it is assumed that the analyst is primarily interested in identifying the causal effect of years of schooling (the treatment), S, on current earnings (the outcome), Y. As such, this representation implies at least four potential causal pathways that we need to control for in order to identify the causal effect of schooling on earnings. The pathways are:

1. $X \rightarrow I \rightarrow S \rightarrow Y$;
2. $X \rightarrow I \rightarrow Y$;
3. $X \rightarrow I \leftarrow P \rightarrow S \rightarrow Y$; and,
4. $X \rightarrow U \rightarrow Y$.

This particular DAG structure implies that to identify the causal effect of years of schooling on earnings, four sufficient conditions (i.e., the graphical identification criteria) must be satisfied for the backdoor criterion:

1. $Y \perp\!\!\!\perp I \mid S, X$;
2. $Y \perp\!\!\!\perp P \mid S, X$;
3. $S \perp\!\!\!\perp X \mid I, P$; and,
4. $X \perp\!\!\!\perp P$.

The first condition implies that current earnings should be independent of parents' income after conditioning on years of schooling and ancestral heritage. The second condition implies that earnings are independent of parents' education after conditioning on years of schooling and ancestral heritage. The third posits that the variable for years of schooling is independent of ancestral heritage after conditioning on parents' income and parents' education. The final condition is that ancestral heritage is independent of parents' education. If any of these hypotheses are strongly refuted (or outright rejected perhaps by prior research), then causal identification is rendered suspect, and the proposed model needs to be revised or re-specified (Barenboim & Hünermund, 2019).

3.4 DAGitty

Only Pearl's (1995) "do-calculus" provides the complete set of criteria for non-parametric causal identification, as outlined above. A discussion of do-calculus is beyond the scope of the current paper, so the reader is referred to the source for exposure to Pearl's (1995) complete criteria.

It can be challenging to learn and defend the graphical identification criteria, but, fortunately, there is an open-source tool that automates the process for researchers. The website for DAGitty (2021) describes the tool as a "...browser-based environment for creating, editing, and analyzing causal model...[the] focus is on the use of causal diagrams for minimizing bias in empirical studies..." A version of DAGitty is also available as a package with R (Textor & van der Zander, 2016). DAGitty, illustrates the causal (and non-causal) pathways in pictorial form (specified by the user), and the program generates the sufficient conditions (contestable hypotheses) based on the analyst's specified model.

4 An Example of the Front-door Criterion using a Monte Carlo Experiment

In the economics literature, the estimated causal parameter of interest is the treatment effect. As outlined above, the principal econometric problem in the estimation of treatment effects (with observational data) is due to unobserved confounding. In this section, we will review a hypothetical returns-to-schooling example that addresses unobserved confounding and the accurate estimation of the average treatment effect.

Through the use of graphical causal models, we discussed different approaches, including instrumental variables, the backdoor criterion, and front-door criterion to dealing with confounding biases. We now turn our attention to a hypothetical returns-to-schooling example by utilizing a Monte Carlo experiment. As before, consider the returns-to-schooling DAG in Figure 8a. The variables are defined as before: X denotes social origin or ancestral heritage; I and P are parent's income and education; S denotes child's years of schooling; and, Y is child's current income or wages.

There is a new variable M that denotes an exogenous mediator variable that lies on the path between years of schooling and current income. As mentioned previously, an example of a mediator is a training program, but this variable may some other type of mechanism that occurs after an individual's schooling and before his or her report of current earnings. Consistent with our discussion above, there is an unobserved confounder U that affects years of schooling and current income – this confounder can represent the unobserved or latent abilities of the units of observation. In this case, U is a post-treatment confounder. If the confounder cannot be adjusted for, then a causal estimate of S on Y cannot be identified. In section two above, we discussed an instrumental variables approach given an exogenous variable Z (absent from the current figure), which is correlated with years of schooling, but independent of the unobserved confounder (ability bias).

As the IV approach has been extensively covered in the economics literature, we will instead focus on the front-door criterion as a means to identify and estimate the causal effect (average treatment effect) of schooling on earnings. To satisfy the FD, recall that the DAG has to satisfy three criteria: (i) M intercepts all directed paths between S and Y; (ii) there is no backdoor path from S to M; and, (iii) all backdoor paths from S to Y are blocked by M. As illustrated in Figure 8a, these conditions are satisfied.

To see how the FD can be applied to the current example, consider Figures 8b and 8c. Under this framework, recall that the average treatment effect of schooling on earnings can be estimated by applying a two-step estimation procedure. In this case, the first is a regression of M on S: $E[M \mid S = s,\, P = p,\, I = i,\, X = x] = \alpha_0 + \alpha_1 s + \alpha_2 p + \alpha_3 i + \alpha_4 x$. The causal paths are represented by the red arcs in Figure 8b. The careful reader may notice that this two-step approach is a means to address the unobserved confounder, which creates collider bias in the full example. By estimating the effects with two separate equations, the confounder is no longer a collider. This is essentially an example of applying the backdoor criterion twice – once for the first model in Figure 8b and again for the second model in Figure 8c.

The second step is a regression of Y on M: $E[Y \mid M = m,\, S = s,\, I = i,\, P = p,\, X = x] = \theta_0 + \theta_1 m + \theta_2 s + \theta_3 i + \theta_4 p + \theta_5 x$. Assuming the DAG conforms to the causal story, then the product of the estimates $\widehat{\alpha_1 \cdot \theta_1}$ should identify the average treatment effect of years of schooling on earnings. The outline and replication code for the backdoor criterion for each of the two models is offered in the online at https://github.com/burnettwesley/graphical_causal_modeling.[8]

4.1 Monte Carlo Example of Returns to Schooling

Given the specification in this section, we illustrate estimation of the average treatment effect (ATE) by simulating the returns to schooling data using Monte Carlo simulations. Specifically, we assume the variables are defined according to the following distributions:

$$
\begin{aligned}
&U, X, P, I \sim N(\mu_j, \sigma_j), \quad j = \{u, x, p, i\} \\
&\qquad S = \pi_0 + \pi_1 U + \varepsilon_s, \\
&\qquad M = \gamma_0 + \gamma_1 S + \gamma_2 P + \gamma_3 I + \gamma_4 X + \varepsilon_m, \\
&\qquad Y = \beta_0 + \beta_1 M + \beta_2 S + \beta_3 I + \beta_4 P + \beta_5 X + \beta_6 U + \varepsilon_y, \text{ and} \\
&\varepsilon_s, \varepsilon_m, \varepsilon_y \sim N(0, 1).
\end{aligned}
$$

Given this specification, the average treatment effect can be recovered from the product of $\widehat{\gamma_1 \cdot \beta_1}$. Notice that the unobserved confounder directly affects years of schooling and current earnings as assumed above.

The distributional assumptions above allow us to estimate three separate effects. The first effect is the estimate of the average treatment based on the front-door criterion. The second effect is the true causal effect of years of schooling on earnings – we estimate the true effect by assuming that we observe U: $E[Y \mid S,\, U]$. The final effect is a biased (naive) estimate of schooling on earnings by ignoring the

[8] The code for the dagitty analysis and Monte Carlo simulations was written and executed in R version 4.1.0.

collider bias in U: $E[Y \mid S]$.

The full calibration of the simulated model, including all of the code for replication, is provided online at https://github.com/burnettwesley/graphical_causal_modeling. For the results below, we assumed the population size consisted of one thousand individuals, and we ran the models with five thousand replications.[9] Based on the Monte Carlo simulations, the distributions of the three effects estimates are provided in Figure 9.

The ATE, based on the front-door criterion, yielded an estimate of approximately 0.77, which implies that, on average, an additional year of schooling would generate about two more dollars per hour in wage ($exp(0.78) \approx 2.15$). In comparison, the true causal effect is estimated at approximately 0.77. The true causal and FD estimates were nearly identical – that is, they only differed at the fourth decimal point. The naive regression, on the other hand, is upwardly biased (as expected) with an estimate of about 0.82, implying about $2.70 more per hour in wage earnings. The difference in estimates seems quite small, which is in part due to the simulated nature of the data. In reality, the unobserved confounder is likely far noisier (and more correlated with the schooling and earnings) than our assumed example. Nonetheless, the naive estimate is biased upward by over six-and-a-half percent.

The discriminating reader may ask if the front-door criterion is an improvement over the instrumental variables approach. After all, the IV literature is very mature, whereas there are only scant examples of FD approach within the economics litera-

[9] As a sensitivity analysis, we ran additional simulations with fewer assumed observations and a smaller number of replications. In general, the results are nearly identical. The FD estimates were nearly identical to the true causal effect, whereas a naive estimate of the treatment on the outcome was always more severely biased.

ture (Deuchert *et al.* , 2019; Bellemare *et al.* , 2020). The answer as to whether the FD offers an improvement over IV depends on the empirical question. As indicated above, IV can be problematic if the instruments are weak (Bound *et al.* , 1995). In the case of FD, an instrument is not needed, but identification relies upon the assumption of the mediator being independent of any unobserved confounds. At a minimum, the front-door criterion provides an alternative estimation strategy that can aid in causal identification.

5 Discussion

In this paper, we discussed the link between causal analysis in economics and the use of graphical causal models. More specifically, we outlined the method of directed acyclic graphs including the terminology and graphical identification criteria.

Graphical causal models offer numerous benefits for applied economic research. Namely, this approach provides an intuitively appealing visual depiction of models as causal chains (Heckman & Pinto, 2015). Put differently, this depiction offers a quick snapshot of the model's assumptions and the causal relationships among the variables. As a complementary tool to regression analysis, DAGs can help guide causal identification and inference, which corresponds to the potential outcomes framework. The DAG approach encodes all of the independence and conditional independence constraints within a model. This encoding of the constraints are offered freely by the open source and automated tool DAGitty, which if used properly can help in the construction and interpretation of DAGs.

The social norms within applied economic research are often to claim ignorance about a causal relationship, and therefore to control for any potential confounding factors as explanatory variables. Contrary to these norms, DAGs help to explain how causal identification is possible within a subset of the control variables, which corresponds to reducing over-control bias. DAGs are beneficial in that they do not generate or place restrictions on the functional form or parametric specifications of models. Finally, DAGs can aid in teaching causal inference to beginners, as the methodology focuses on identification rather than statistical estimation and inference.

As discussed in the introduction, however, DAGs should not necessarily replace or supersede the potential outcomes framework as applied to economic research (Imbens, 2020). Further, DAGs should not be construed as a one-stop solution for applied research. An analyst must still bring prior knowledge to a particular research question including a fundamental understanding of: economic theory, regression analysis (econometrics), knowledge of the past literature, and experience based on past research. Moreover, DAGs can aid in identifying relations among variables but that does not necessarily imply that causal inference is warranted. In other words, the analyst's causal story has to conform to the graphical depiction; otherwise, any claims of causal inference are suspect.

Heckman & Pinto (2015) validly criticized the inability of DAGs to accommodate nonrecursive simultaneous equations models, which is still a topic of on-going research (Spirtes, 1993; Richardson, 1994; Koster, 1996). Imbens (2020) also makes a valid argument that the potential outcomes framework is better suited to economic theory by more effectively incorporating restrictions beyond the conditional independence constraints. Nevertheless, DAGs can still help aid, at least qualitatively, in an analyst's attempt at identifying cause-and-effect relationships. To this end, DAGs are a complementary tool, which will help push the credibility and causation revolutions forward into the future.

References

Angrist, J.D., & Krueger, A.B. 1991. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106**(4), 979–1014.

Angrist, J.D., & Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, New Jersey: Princeton University Press.

Barenboim, E., & Hünermund, P. 2019. *Causal inference and data-fusion in econometrics*. Accessed online January 2020, at `https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=EEAESEM2019&paper_id=491`.

Baron, R.M., & Kenny, D.A. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173–1182.

Bellemare, M.F., Bloem, J.R., & Wexlar, N. 2020. *The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion*. Accessed online January 2021, at https://www.canr.msu.edu/afre/events/Bellemare%20and%20Bloem%20(2020).pdf.

Bound, J., Jaeger, D.A., & Baker, R.M. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, **90**, 443–450.

Carney, R.N., & Levin, J.R. 2002. Pictorial illustrations still improve students' learning from texts. *Educational Psychology Review*, **14**, 5–26.

Cinelli, C., Forney, A., & Pearl, J. 2021. *A crash course in good and bad controls*. Technical Report R-493, accessed online May 2021, https://ftp.cs.ucla.edu/pub/stat_ser/r493.pdf.

Clay, K., Lingwall, J., & Stephens Jr, M. 2021. Laws, educational outcomes, and returns to schooling evidence from the first wave of U.S. state compulsory attendance laws. *Labour Economics*, **68**, 1–10.

Corak, M. 2013. Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, **27**(3), 79–102.

Cunningham, S. 2021. *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press.

DAGitty. 2021. *Welcome to DAGitty!* Accessed online March 2021 at `www.dagitty.net`.

Deuchert, E., Huber, M., & Schelker, M. 2019. Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery. *Journal of Business and Economic Statistics*, **37**(4), 710–720.

Elwert, F. 2013. Graphical causal models. *Pages 245–273 of:* Morgan, S.L. (ed), *Handbook of causal analysis for social research*. Dordrecht, Norway: Springer Science + Business Media.

Elwert, F. 2019. *Lecture notes in Directed Acyclic Graphs for Causal Inference*.

Elwert, F., & Winship, C. 2002. Commentary: Population versus individual level causal effects. *International Journal of Epidemiology*, **31**, 432–434.

Elwert, F., & Winship, C. 2014. Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology*, **40**, 31–53.

Feigenbaum, J.J., & Tan, H.R. 2019. *The return to education in the mid-20th century: Evidence from twins*. NBER Working Paper Number 26407.

Goldin, C., & Katz, L. 2000. Education and income in the early 20th century: Evidence from the prairies. *Journal of Economic History*, **60**, 782–818.

Grätz, M. 2019. *When less conditioning provides better estimates: Overcontrol AND collider bias in research on intergenerational mobility*. Swedish Institute for Social Research, Stockholm University, Working Paper 02/2019, https://www.su.se/polopoly_fs/1.433943.1554544422!/menu/standard/file/Graetz2019.pdf.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. 2002. *A distribution-free theory of nonparametric regression*. New York, NY: Springer-Verlag.

Heckman, J., & Pinto, R. 2015. Causal analysis after Haavelmo. *Econometric Theory*, **31**(1), 115–151.

Holland, P.W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Imbens, G.W. 2020. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, **58**(4), 1129–1179.

Koster, J.T.A. 1996. Markov properties of nonrecursive causal models. *Annals of Statistics*, **24**(5), 2148–2177.

Manski, C.F. 2007. *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

Mincer, J. 1958. Investment in human capital and personal income distribution. *Journal of Political Economy*, **66**(4), 281–302.

Mincer, J. 1974. *Schooling, experience and earnings*. New York, NY: Columbia University Press.

Morgan, S.L., & Winship, C. 2015. *Counterfactuals and Causal Inference*. 2 edn. New York, NY: Cambridge University Press.

Neyman, J. 1923 [1990]. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**(4), 465–472. Trans. D.M. Dabrowska and T.P. Speed.

Oreopoulos, P., & Salvanes, K. 2011. Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, **25**(1), 159–184.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. 2 edn. Oxford, UK: Morgan Kaufmann Publishers, Inc.

Pearl, J. 1993. Comment: Graphical models, causality and intervention. *Statistical Science*, **8**(3), 266–269.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, **82**(4), 669–710.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Pearl, J. 2009. *Causality*. 2 edn. New York, NY: Cambridge University Press.

Pearl, J. 2018. *Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Pearl, J., Glymour, M., & Jewell, N.P. 2016. *Causal inference in statistics: A primer*. West Sussex, UK: John Wiley & Son Ltd.

Richardson, T. 1994. Equivalence in non-recursive structural equation models. *In:* Dutter, R., & Grossmann, W. (eds), *Compstat: Proceeding in Computation Statistics*. Berlin, Germany: Springer-Verlag.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

Rubin, D.B. 1980. Comment on 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by Basu. *Journal of the American Statistical Association*, **75**, 591–593.

Rubin, D.B. 1986. Which ifs have causal answers (Comment on 'Statistical and Causal Inference' by Paul W. Holland). *Journal of the American Statistical Association*, **81**, 961–962.

Rubin, D.B. 2006. *Matched Sampling for Causal Effects*. Cambridge, England: Cambridge University Press.

Shpitser, I., & VanderWeele, T.J. 2011. A complete graphical criterion for the adjustment formula in mediation analysis. *International Journal of Biostatistics*, **7**(1), 1–24.

Spirtes, P. 1993. *Directed cyclic graphs, conditional independence, and non-*

recursive linear structural equation models. Department of Philosophy Technical Report CMU-Phil-35, Carnegie Mellon University.

Spirtes, P. 2005. Graphical models, causal inference, and econometric models. *Journal Economic Methodology*, **12**(1), 1–33.

Staiger, D., & Stock, J.H. 1997. Instrumental variables regression with weak instruments. *Econometrica*, **65**, 557–586.

Stephens Jr, M., & Yang, D.-Y. 2014. Compulsory education and the benefits of schooling. *American Economic Review*, **104**(6), 1777–1792.

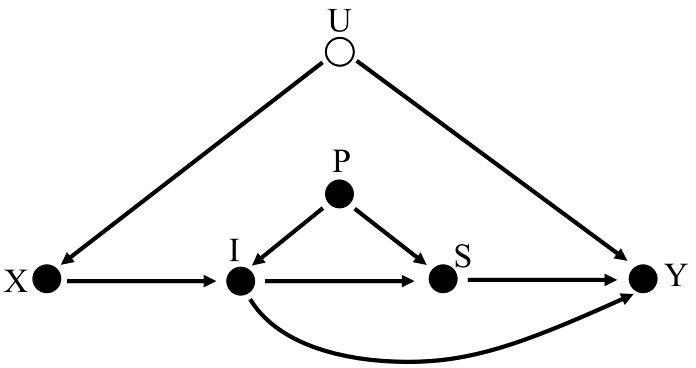
Textor, J., & van der Zander, B. 2016. *Dagitty: Graphical analysis of structural causal models*. Accessed online October 2019 at `https://cran.r-project.org/web/packages/dagitty/index.html`.

Acknowledgements

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. This research was supported by the intramural research programs of the USDA Economic Research Service and the College of Charleston.

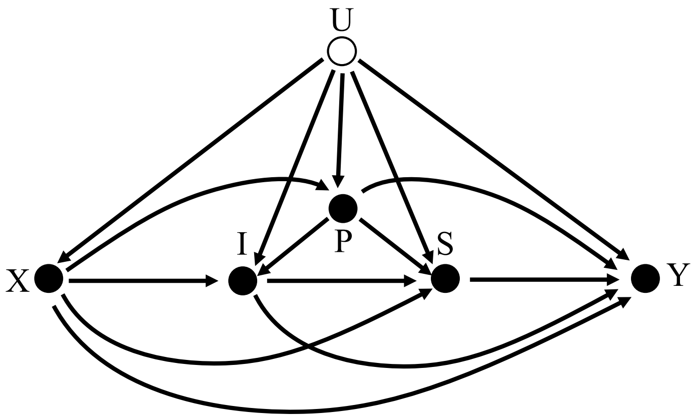
6 Figures

Figure 1: Directed acyclic graph representing returns-to-schooling model



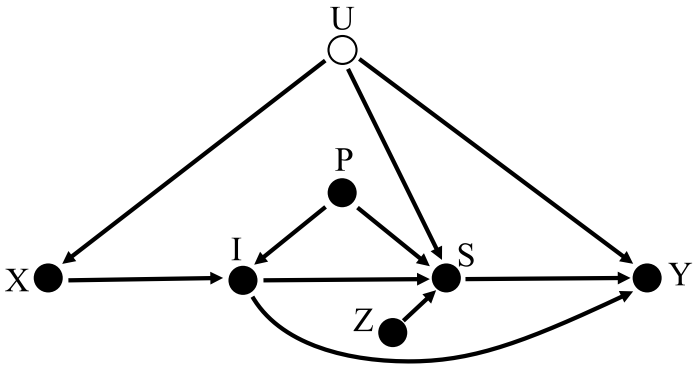
Notes: X = ancestral heritage; I = parents' income; P = parents' education; S = years of schooling (treatment); Y = earnings (outcome); and, U = unobserved confounder.

Figure 2: Complete directed acyclic graph



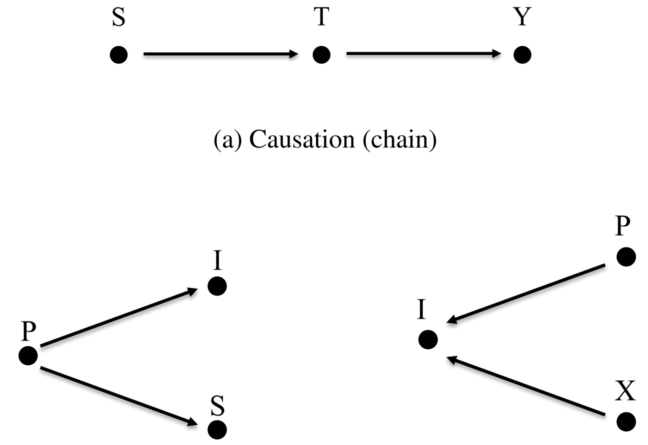
Notes: X = ancestral heritage; I = parents' income; P = parents' education; S = years of schooling (treatment); Y = earnings (outcome); and, U = unobserved confounder.

Figure 3: Directed acyclic graph with latent “ability" confounding years of schooling



Notes: X = ancestral heritage; I = parents’ income; P = parents’ education; S = years of schooling (treatment); Y = earnings (outcome); and, U = unobserved confounder.

Figure 4: Causation, Confounding, and Endogenous Selection



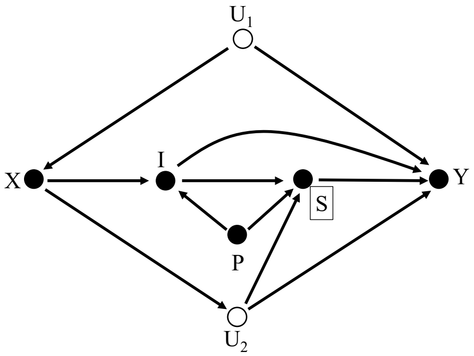
(a) Causation (chain)

(b) Confounding (fork)

(c) Endogenous selection (inverted fork)

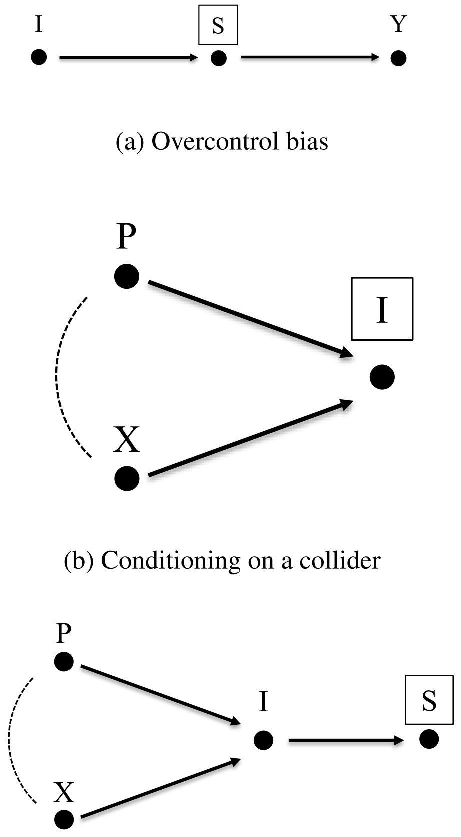
Notes: X = ancestral heritage; I = parents’ income; P = parents’ education; S = years of schooling; Y = earnings; T = work training program; and, U = unobserved confounder.

Figure 5: Over-control bias introduced by conditioning on child's education



Notes: X = ancestral heritage; I = parents' income; P = parents' education; S = years of schooling (treatment); Y = earnings (outcome); and, U_i = unobserved confounders for $i = 1, 2$.

Figure 6: Over control bias and endogenous selection



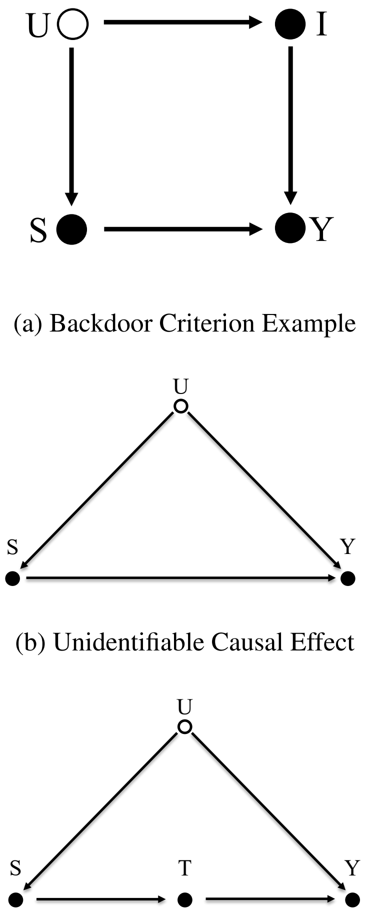
(a) Overcontrol bias

(b) Conditioning on a collider

(c) Conditioning on the descendent of a collider

Notes: X = ancestral heritage; I = parents' income; P = parents' education; S = years of schooling (treatment); Y = earnings (outcome); and, U = unobserved confounder.

Figure 7: Examples of Backdoor and front-door Criterion



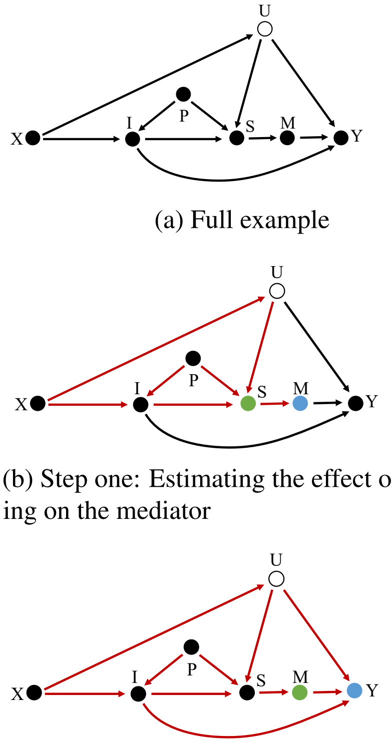
(a) Backdoor Criterion Example

(b) Unidentifiable Causal Effect

(c) Front-door Criterion Example

Notes: I = parents' income; S = years of schooling (treatment); Y = earnings (outcome); T = work training program; and, U = unobserved confounder.

Figure 8: Returns-to-Schooling Example Using the Front-door Criterion



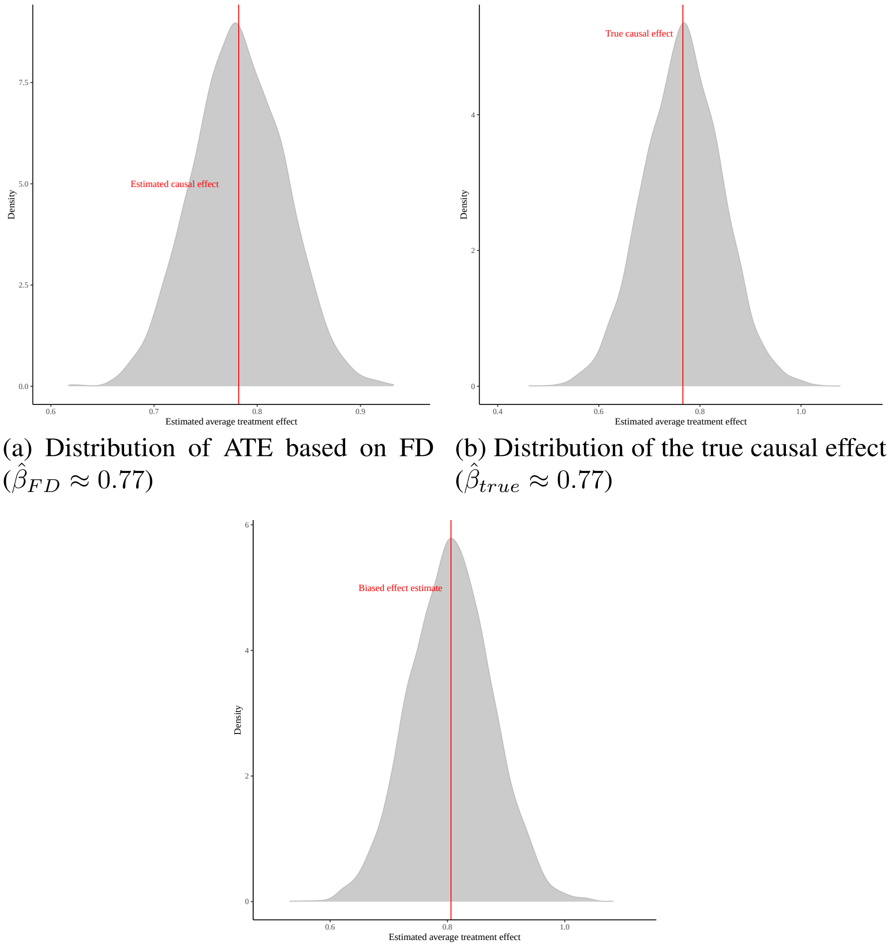
(a) Full example

(b) Step one: Estimating the effect of schooling on the mediator

(c) Step two: Estimating the effect of the mediator on earnings

Notes: X = social origin; I = parents' income; P = parents' education; S = years of schooling (treatment); M = exogenous mediator (e.g., work training program); U = unobserved confounder; and, Y = earnings (outcome). Red arcs represent conditioning paths for each step of two-stage estimation procedure. The green (blue) circle represents the treatment (outcome) variable.

Figure 9: Causal Effects Estimates Based on Monte Carlo Simulations



(a) Distribution of ATE based on FD ($\hat{\beta}_{FD} \approx 0.77$)

(b) Distribution of the true causal effect ($\hat{\beta}_{true} \approx 0.77$)

(c) Distribution of the biased ATE ($\hat{\beta}_{bias} \approx 0.82$)

Online Appendix

A Potential Outcomes

The potential outcomes framework is often referred to as the Neyman-Rubin model. This approach originated with the work of Neyman (1923 [1990]), who developed a nonparametric model where each observational unit has two potential outcomes, but only one outcome is observed.[10] If the unit is treated then it is assigned a binary value of unity, or a unit is assigned a value of zero if untreated.[11]

Rubin (1974, 2006) extended the model into a general framework for causal inference for observational research. Moreover, Holland (1986) wrote a review article that highlighted some of the philosophical implications of the framework. Consequently, this approach is referred to as the Neyman-Rubin model or sometimes the Neyman-Rubin-Holland model of causal inference. For ease of exposition, we will refer to the potential outcomes model as the Neyman-Rubin model throughout the remainder of the paper.

To fix ideas, consider the binary treatment $T_i = \{1, 0\}$ for individual i, where $T_i = 1$ indicates "treatment" and $T_i = 0$ indicates "control." Each unit of observation has two potential outcomes, $Y_i^T = \{Y_i^1, Y_i^0\}$, one for each value of the treatment. A potential outcome is the outcome that would be realized if an observational unit received a specific value of the treatment, where Y^1 is the potential outcome if *i* received the treatment, and Y^0 is the potential outcome if *i* did not

[10] The analysis can be extended to more than two potential outcomes, but for simplicity we assume only two.

[11] We discuss binary outcomes here for convenience; however, this framework works just as well for continuous variables.

receive the treatment.

Observable outcomes, Y_i, are distinct from potential outcomes in that potential outcomes are hypothetical random variables that differ across the population, whereas observables are factual random variables (Cunningham, 2021). Putting the treatment indicator and the potential outcomes together, we can define each unit's observable outcome according to a switching model:

$$Y_i = T_i\, Y_i^1 + (1 - T_i)\, Y_i^0. \tag{1}$$

The switching model reveals the fundamental problem of causal inference, which is that only one state (or outcome) of the world can be observed and estimated.

A causal effect (or treatment effect) is defined as the difference between the two potential outcomes (or two different states of the world). Elwert & Winship (2002) refer to this difference as the “individual-level causal effect" or (ICE):

$$\text{ICE} = \delta_i = Y_i^1 - Y_i^0. \tag{2}$$

The ICE model answers the question of what would have happened to an observational unit if exposed to the treatment rather than the control. The ICE specification offers a simple estimate of treatment, but according to the switching model, we only observe one state of the world, so we cannot necessarily calculate the treatment effect. Even if we could estimate the ICE model, it would be biased if the treatment effect differs across observed individuals (effect heterogeneity).

Thus, it is often more useful to measure the average causal effect or average treatment effect (ATE), which is the population average of the individual-level causal effects:

$$\text{ATE} = E[\delta_i] = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]. \tag{3}$$

As before, the true average treatment effect is unknowable because one can never observe both potential outcomes for a given individual unit i (Holland, 1986).

In an attempt to determine the difference between potential outcomes, we can instead examine a naïve comparison of averages between treated and untreated observational units. Morgan & Winship (2015) refer to the *estimated* ATE as the "naïve average treatment effect." Following Angrist & Pischke (2009), the comparison of averages is formally linked to the average treatment effect by the equation:

$$\begin{aligned}\underbrace{E[Y_i|T_i = 1] - E[Y_i|T_i = 0]}_{\text{Observed difference in average treatment outcome}} &= \underbrace{E[Y_i^1|T_i = 1] - E[Y_i^0|T_i = 1]}_{\text{Average treatment effect on the treated}} \\ &+ \underbrace{E[Y_i^0|T_i = 1] - E[Y_i^0|T_i = 0]}_{\text{Selection bias}}. \end{aligned} \tag{4}$$

Notice that the term on the left-hand sign of the equation is the observed difference in the average treatment outcome; hence, the superscripts are omitted.

The first term on the right-hand side of equation (4) is the average population treatment effect for the groups of units that have been assigned to the treatment (or the average treatment effect on the treated (ATT)). It captures the average difference between the treated group $E[Y_i^1|T_i = 1]$ and what would have happened to the same group had they not received the treatment $E[Y_i^0|T_i = 1]$.

The second term on the right-hand side of equation (4) is the selection bias.

The second term is sometimes referred to as the average treatment effect on the controlled (ATC). It is the average difference between what would have happened to the control group if they had received treatment $E[Y_i^0|T_i = 1]$ and the untreated, control group $E[Y_i^0|T_i = 0]$. As an example, Angrist & Pischke (2009) describe Y_i^0 as the health status of a person within the control group and T_i as the treatment of hospitalization. Sick people are more likely than the healthy to seek treatment, so people who are hospitalized have worse values of Y_i^0 (i.e., the average treatment on the controls is non-negative), which leads to negative bias. If the selection bias is large, then it may mask a treatment effect.

One way to solve the selection problem is through random assignment of treatment, which leads to independence between the treatment and the control groups:

$$\begin{aligned} E[Y_i|T_i = 1] - E[Y_i|T_i = 0] &= E[Y_i^1|T_i = 1] - E[Y_i^0|T_i = 0] \\ &= E[Y_i^1|T_i = 1] - E[Y_i^0|T_i = 1] \\ &= E[Y_i^1 - Y_i^0|T_i = 1] \\ &= E[Y_i^1 - Y_i^0]. \end{aligned} \tag{5}$$

Random assignment leads to independence of Y_i^0 and T_i, so the term $E[Y_i^0|T_i = 1]$ can be swapped with $E[Y_i^0|T_i = 0]$ in the second line of equation (5) (Angrist & Pischke, 2009).

Random assignment of units to treatment or control, as in randomized control trials, is the theoretical ideal. However, randomization is often infeasible for social science research. There are workarounds to make causal claims with nonrandomized or observation data, but first we must understand the assumptions for causal inference.

B Assumptions for Causal Inference

In order to draw inference from a causal estimate with observational data, there are three basic assumptions that must be satisfied. In this section, we outline all three of these assumptions.

B.1 Ignorability Assumption

The first of these assumptions is “ignorability."

Definition 3 *(Ignorability assumption) Valid inference of the average treatment effect requires independence between potential outcomes and treatment:*

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp T_i, \forall i. \tag{6}$$

In equation (6), the variable Y_i^1 represents the potential outcome in which a unit or participant receives the treatment; Y_i^0 represents the potential outcome where the unit does not receive the treatment. The variable T_i is generally interpreted as a binary treatment, but it can be generalized to a factor or continuous variable. This equation indicates that the potential outcomes are statistically independent, “$\perp\!\!\!\perp$”, of the treatment. With the ignorability condition satisfied, the treatment and control groups are comparable, and the naïve comparison identifies the average treatment effect:

$$\underbrace{E[Y_i|T_i=1]-E[Y_i|T_i=0]}_{\text{Association}} = \underbrace{E[Y_i^1 - Y_i^0]}_{\text{Causation}}. \tag{7}$$

The ignorability assumption is often satisfied through true randomization (e.g., through direct researcher manipulation) of assignment of treatment and controls. In the absence of randomization, a stronger assumption is needed.

B.2 Conditional Ignorability Assumption and Common Support

The stronger form of the ignorability assumption is referred to as “conditional ignorability" (Holland, 1986).

Definition 4 *(Conditional ignorability assumption) Valid inference of the average treatment effect requires a set of conditioning variables,* $\mathbf{X}$*, which determine assignment to treatment; and, conditioning on* $\mathbf{X}$ *generates independence between the treatment and the potential outcomes:*[12]

$$(Y^1, Y^0) \perp\!\!\!\perp T \mid \mathbf{X}. \tag{8}$$

Equation (8) requires that the potential outcomes, Y^1 and Y^0, are jointly independent of the treatment assignment within groups defined by the values of $\mathbf{X}$.

Conditional ignorability is a stronger assumption than ignorability as the former implies (nearly) perfect stratification of the data (Elwert, 2019). In other words, the potential outcomes are ignorable after conditioning on all of the systematic

[12] Generally speaking, to condition on a variable is to introduce information about a variable into the analysis (Morgan & Winship, 2015).

components of treatment choice. This implies that within each level of $\mathbf{X}$, the treatment effect is effectively randomly assigned.

Given conditional ignorability, the naïve comparison, between the observed outcomes in the treatment and control groups (within levels of $\mathbf{X}$), provides an unbiased estimate of the conditional, average treatment effect $\text{ATE}_x = E[\delta_i | X_i = x]$. The challenge is in determining whether the conditional ignorability assumption holds, which implies no confounding.

It is remarkably difficult to prove that the unconfoundedness condition as it depends on potential outcomes that are not always observed. To help prove that conditional ignorability is credible, the analyst must then demonstrate how the treatment is assigned – that is, the assignment mechanism. In other words, the analyst must bring substantive knowledge (for example, economic theory) to the phenomenon, so that she can propose and defend the necessary assumptions within the model. As an aid to demonstrating conditional ignorability, causal graphical models, including DAGs, provide a understandable set of tools to demonstrate a credible assignment mechanism.

B.3 Stable Unit Treatment Value Assumption

The final assumption for causal inference is referred to as the stable unit treatment value assumption or SUTVA (Rubin, 1980). SUTVA is often referred to as the “no spillover" assumption. This assumption, in part, addresses potential forms of treatment-effect-heterogeneity bias.

Definition 5 *(SUTVA) A prior assumption that requires that the response of a particular unit depends only on the treatment to which he or she was assigned, not* *the treatment others receive within the experiment (Rubin, 1986, p. 961).*

If both the (conditional) ignorability and SUTVA assumptions are satisfied, then a researcher has a strong foundation to plausibly draw causal inference. In the next section, we will discuss how graphical causal models can help to elucidate and satisfy these two assumptions with observational data.

C Additional Terminology

In addition to the structure of DAGs, there is complementary terminology regarding the graph's nodes (Elwert, 2019). A "descendant" node is defined as all nodes that directly or indirectly cause the node. Referring to back Figure 1, S and Y are descendants of X or desc$(X) = \{S, Y\}$. The "children" of a node are defined as all nodes directly caused by a preceding node. For example, in Figure 1 I is a child of X or child$(X) = \{I\}$. "Ancestors" are nodes directly or indirectly causing another node. For example, the ancestors of S are I, and X or an$(S) = \{I, X\}$. Finally, "parents" are all direct causes of the node or pa$(I) = \{X\}$.