Theoretical Background

Causility, Associations, and (In)dependence

Graphical Causal Models

Graphical models are the easiest way to conceptualize causal systems. Pioneered by XXXX and YYYY, they allow to visualize causal relationships, which eases development and understanding of causal models.

A graphical causal model visualizes the exposure, outcome, covariates, and their (assumed) causal relationship. In the following, we will usually note the exposure (or independent variable) with X, the outcome (or dependent variable) with Y, and covariates (all other variables) with other letters. Note, that depending on the research question, the assignment of exposure, outcome, and covariates may change within the same model, so there can be instances, where the conventional naming cannot be followed.

Variables in graphical causal model are connected by arrows. An arrow between X and Y means, that a direct causal relationship between the two is assumed (see Figure 1) . The direction of the arrow tells the direction of causality. As in Figure 1, $X \rightarrow Y$ means that X causes Y (and not the other way around). For our definition of causality this means, if we intervene on X we expect to see a change in Y.

The direction of causality has to be determined by theoretical knowledge. It cannot be found in the data itself. Suppose that in our first example in Figure 1, X is biological sex and Y is endurance performance. It appears obvious, that a causal relationship between both exist (though it is certainly much more complicated than that seen in the simple model). However, the fact that it is sex that causes performance — and not the other way around — is based purely on theoretical knowledge and understanding of the world. There are neither randomized trials for proof (because you cannot randomly assign sex), nor controlled interventions (because you cannot easily intervene on sex) possible. Ultimately, the direction of causality is an assumption by the researcher.

Causal systems in the world are usually more complex than consisting of only exposure and outcome, and so are the graphical causal models depicting them. A more complex graph is displayed in Figure 2. X and Y are not directly connected anymore, but indirectly via B. This is called a *causal path*. We will later see, that some models also have non-causal paths.

The graph in Figure 2 is called an directed acyclic graph (DAG). It is directed, because all paths have arrows (the direction of causality is set). It is acyclic, because there are no circular paths in it. Finally, it is a graph. All graphs in this thesis will be DAGs, because the presented tools work only for these, and most research problems can be adequately formulated using them.

What is inside a DAG is as important as what is not inside it. A DAG should depict all causal relations important to the research question. If two variables are not connected, we assume

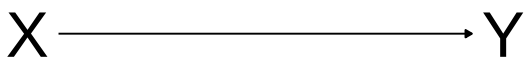

Figure 1: A simple graphical causal model with two variables. The variable X (exposure) is assumed to cause the variable Y (outcome). No other variables are believed to influence this process.

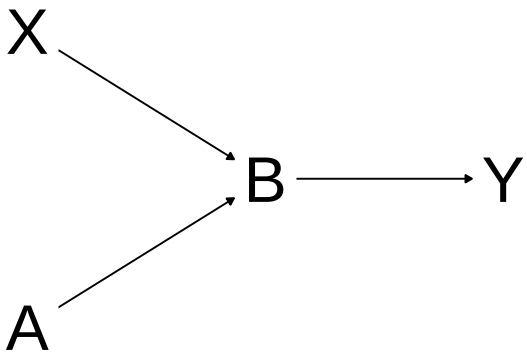

Figure 2: A more complex graphical causal model with four variables. X and A both cause B, which in turn causes Y.

that they do not causally relate to each other. For example, in Figure 2, there is no direct connection between X and A, or between X and Y.

DAGs tell a story. For example, we can assign the variables in Figure 2 to a very simple model of endurance performance. Let X be the biological sex, A the nutrition status, B the physiological capacity to perform endurance tasks, and Y the endurance performance in a competition. Our model assumes, that sex and nutrition both directly affect the physiological capacity, and this in turn affects the performance. On the other hand, it assumes that sex and nutrition are not causally related, and that both sex and nutrition have no direct effect on performance, but only an indirect effect via physiological capacity.

Error Terms in Causal Modeling

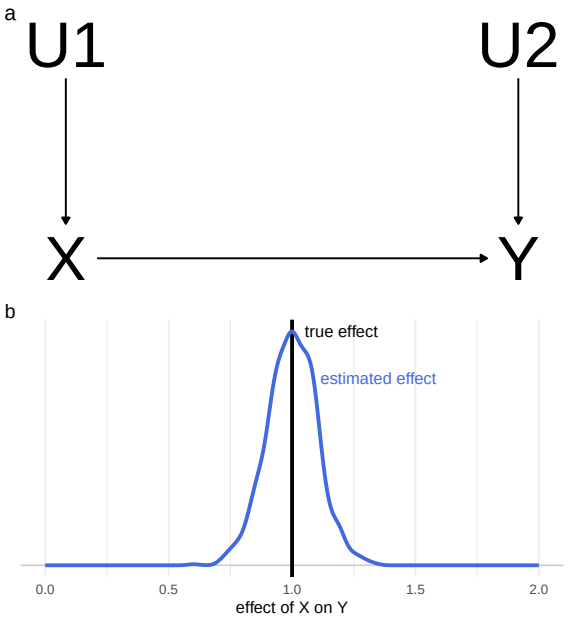

Figure 3: A simple causal path, with random error. (a) X causes Y, but both variables are influenced by other unobserved variables (random error). (b) A simulation of the model. The random error adds uncertainty to the estimate of the causal effect, but no bias.

Modeling Causal Systems

DAGs can be understood as linear models.

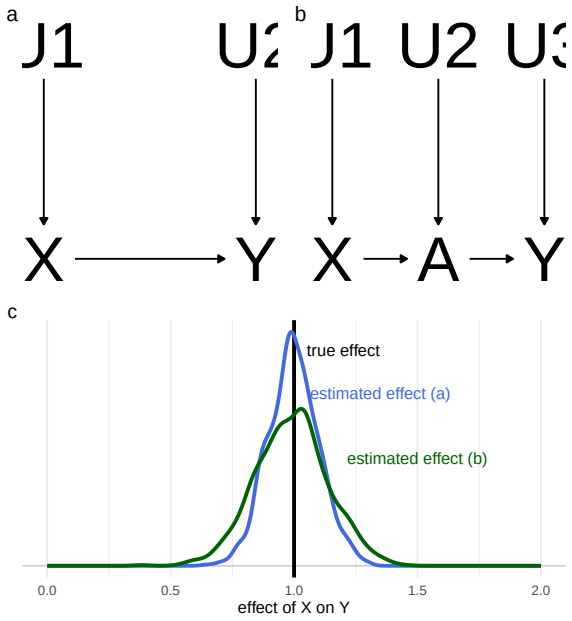

Figure 4: Random errors in a causal path. (a) X causes Y directly. Both variables are influenced by random errors. (b) X causes Y via A. All three variables are influenced by random errors. (b) A simulation of the effect of X on Y in both models. The chain creates uncertainty in the effect estimate, but no bias.

DAGs are an abstract concept to describe research problems. This level of abstraction allows to plan a study and its data analysis on a conceptual level. For the actual data analysis, a DAG has to be filled with data and functions.

Arrows in a DAG represent a causal relationship, but they do not specify the type of this relationship. $X \rightarrow Y$ can mean a strong linear affect of X on Y, or a weak negative curvelinear relation. While the researcher is free to choose what exact type of relation a DAG assumes for a set of two variables, a linear relationship is the common choice.

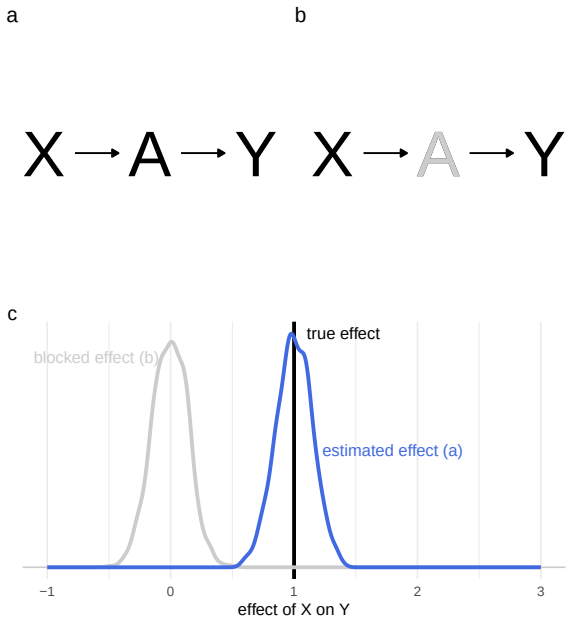

Figure 5: A causal path blocked by conditioning. (a) X causes Y via A. (b) The causal path is blocked, because the analysis conditions on A. As all affects of X on Y trail through A, no causal effect remains. (c) A simulation of the effect of X on Y in both models. Blocking removes the true causal effect entirely.

Confounders and Colliders

Conditioning Rules: The Backdoor Criterion

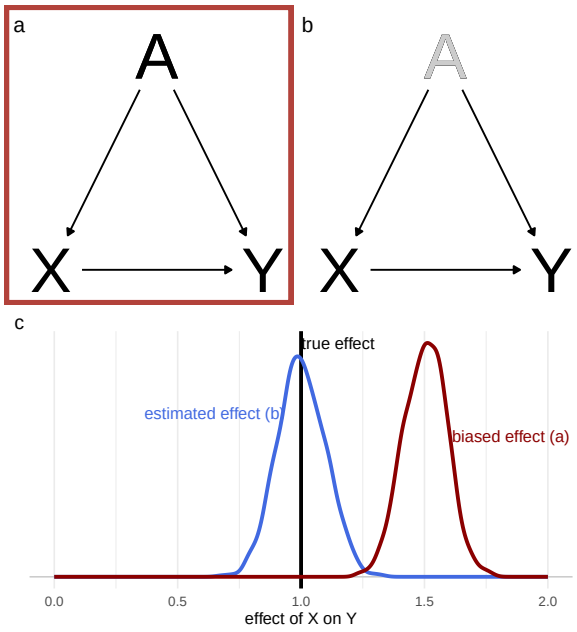

Figure 6: A graphical example of confounding. Both X and Y share a common cause A. (a) This confounder biases determining the causal effect of X on Y. (b) Conditioning on A removes the bias in the analysis. (c) A simulation of the effect of X on Y in both models.

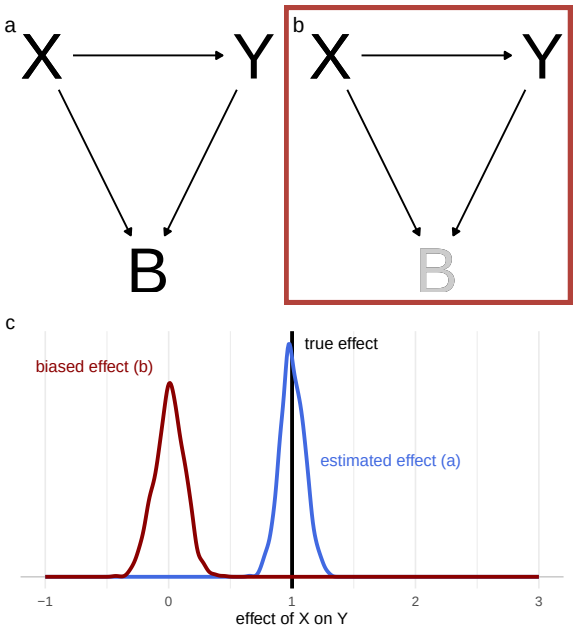

Figure 7: A graphical example of collider bias. Both X and Y directly affect the collider B. (a) As long as B is not conditioned on, the causal effect of X on Y is unbiased. (b) Conditioning on B will introduce bias in the model. (c) A simulation of the effect of X on Y in both models.