Chapter 2

Causal models

Causal inference aims to infer the outcome of an intervention in some system from data obtained by observing (but not intervening in) it. To do this we need terminology to describe actions and how we anticipate the system should respond to them. Three key approaches have emerged; counterfactuals, structural equation models and causal Bayesian networks. In this chapter we will examine the problems these approaches allow us to solve, the assumptions they rely on and how they differ. We will also use them to describe the following simplified examples. The aim is to demonstrate the notations and formalisms needed to tackle more interesting problems later on.

Example 1. Suppose a pharmaceutical company wants to assess the effectiveness of a new drug on recovery from a given illness. This is typically tested by taking a large group of representative patients and randomly assigning half of them to a treatment group (who recieve the drug) and the other half to a control group (who recive a placebo). The goal is to determine the clinical impacts of the drug by comparing the differences between the outcomes for the two groups (in this case, simplified to only two outcomes - recovery or non-recovery). We will use the variable X (1 = drug, 0 = placebo) to represent the treatment each person receives and Y (1 = recover, 0 = not recover) to describe the outcome.

Example 2. Suppose we want to estimate the impact on high school graduation rates of compusory preschool for all 4 year olds. We have a large cross-sectional dataset on a group of 20 year olds that records if they attended pre-school, if they graduated high school and their parents socio-economic status (SES). We will let $X \in \{0, 1\}$ indicate if an individual attended pre-school, $Y \in \{0, 1\}$ indicate if they graduated high school and $Z \in \{0, 1\}$ represent if they are from a low or high SES background respectively.[1]

2.1 Causal Bayesian networks

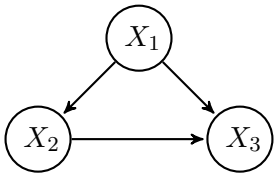
Causal Bayesian networks are an extension of Bayesian networks. A Bayesian network is a graphical way of representing how a distribution factorises. Any joint probability distribution can be factorised into a product of conditional probabilities. There are multiple valid factorisations, corresponding to permutations of variable ordering.

$$P(X_1, X_2, X_3, ...) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)... \tag{2.1}$$

[1]There has been substantial empirical work on the effectevness of early childhood education including a landmark randomised trial, the Perry Preschool project, which ran from 1962-1967 [62].

We can represent this graphically by drawing a network with a node for each variable and adding links from the variables on the right hand side to the variable on the left for each conditional probability distribution, see figure 2.1. If the factorisation simplifies due to conditional independencies between variables, this is reflected by missing edges in the corresponding network. There are multiple valid Bayesian network representations for any probability distribution over more than one variable, see figure 2.2 for an example.

Figure 2.1: A general Bayesian network for the joint distribution over three variables. This network does not encode any conditional independencies between its variables and can thus represent any distribution over three variables.



The statement that a given graph G is a Bayesian network for a distribution P tells us that the distribution can be factorised over the nodes and edges in the graph. There can be no missing edges in G that do not correspond to conditional independencies in P (the converse is not true G can have extra edges). If we let $parents_{X_i}$ represent the set of variables that are parents of the variable X_i in G then we can write the joint distribution as;

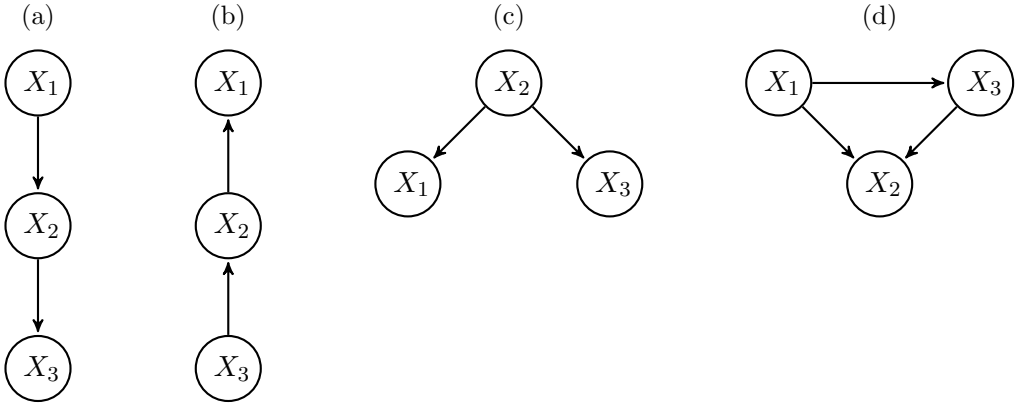
$$P(X_1 ... X_N) = \prod_{i=1...N} P(X_i | parents_{X_i}) \tag{2.2}$$

A causal Bayesian network is a Bayesian network in which it a link $X_i \rightarrow X_j$, by definition, implies X_i causes X_j. This means that if we intervene and change the value of X_i, we expect X_j to change, but if we intervene to change X_j, X_i will not change. We need some notation to describe interventions and represent distributions over variables in the network after an intervention. In this thesis I use the do operator introduced by Pearl [38].

Definition 3. The do-notation

- do(X=x) denotes an intervention that sets the random variable(s) X to x.
- $\mathrm{P}\{Y|do(X)\}$ is the distribution of Y conditional on an *intervention* that sets X. This notation is somewhat overloaded. It may be used represent a probability, a probability

Figure 2.2: Some valid Bayesian networks for a distribution that an be factorised as $\mathrm{P}\{X_1, X_2, X_3\} = \mathrm{P}\{X_1\}\,\mathrm{P}\{X_2\}\,\mathrm{P}\{X_3|X_2\}$ (which implies $X_3 \perp\!\!\!\perp X_1|X_2$)



distribution/mass function or a family of distribution functions depending on if the variables are discrete or continuous and whether or not we are treating them as fixed. For example it could represent

- The probability $\mathrm{P}\{Y = 1|do(X = x)\}$ as a function of x
- The probability mass function for a discrete Y : $\mathrm{P}\{Y|do(X = x)\}$
- The probability density function for a continuous Y : $f_Y(y|do(X = x))$
- a familiarly of density/mass function for Y paramaterised by x.

Where the distinction is important and not clear from context we will use one of the more specific forms above.

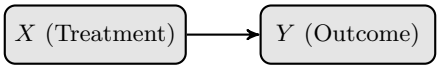
Theorem 4 (Truncated product formula)**.** *If G is a causal network for a distribution P defined over variables $X_1...X_N$, then we can calculate the distribution after an intervention where we set $Z \subset X$ to z, denoted $do(Z = z)$ by dropping the terms for each of the variables in Z from the factorisation given by the network [38].*

$$\mathrm{P}\{X_1...X_N|do(Z = z)\} = \begin{cases} \prod_{i \notin Z} \mathrm{P}\{X_i|parents_{X_i}\} & \text{if } (X_1...X_N) \text{ consistent with } Z = z \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Theorem 4 does not hold for standard Bayesian networks because there are multiple valid networks for the same distribution. The truncated product formula will give different results depending on the selected network. The result is possible with causal Bayesian networks because it follows directly from the assumption that the direction of the link indicates causality. In fact, from the interventionist viewpoint of causality, the truncation product formula defines what it means for a link to be causal.

Returning to example 1, and phrasing our query in terms of interventions; what would the distribution of outcomes look like if everyone was treated $\mathrm{P}\{Y|do(X = 1)\}$, relative to if no one was treated $\mathrm{P}\{Y|do(X = 0)\}$? The treatment X is a potential cause of Y, along with other unobserved variables, such as the age, gender and the disease sub type of the patient. Since X is assigned via deliberate randomisation we know that it is not effected by any latent variables. The causal Bayesian network for this scenario is shown in figure 2.3 .This network represents the (causal) factorisation $\mathrm{P}\{X, Y\} = \mathrm{P}\{X\}\,\mathrm{P}\{Y|X\}$, so from equation (2.3), $\mathrm{P}\{Y|do(X)\} = \mathrm{P}\{Y|X\}$. In this example, the interventional distribution is equivalent to the observational one.

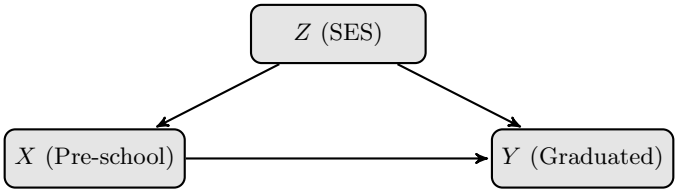
Figure 2.3: Causal Bayesian network for example 1



In example 2 we are interested $\mathrm{P}\{Y|do(X = 1)\}$, the expected high-school graduation rate if we introduce universal preschool. We could compare it to outlawing preschool $\mathrm{P}\{Y|do(X = 0)\}$ or the current status quo $\mathrm{P}\{Y\}$. It seems reasonable to assume that preschool attendence affects the likelihood of high school graduation [2] and that parental socio-economic status would affect *both* the likelihood of preschool attendance and high school graduation. If we assume that socio-economic status is the only such variable (nothing else effects both attendance *and* graduation), we can represent this problem with the causal Bayesian network in figure 2. In this case, the interventional distribution is not equivalent to the observational one. If parents

[2]The effect does not have to be homogenius, it may depend non-linearly on characteristics of the child, familiy and school.

Figure 2.4: Causal bayesian network for example 2



with high socio-economic status are more likely to send their children to preschool and these children are more likely to graduate high school regardless, comparing the graduation rates of those who attended preschool with those who did not will overstate the benefit of preschool. To obtain the interventional distribution we have to estimate the impact of preschool on high school graduation for each socio-economic level separately and then weight the results by the proportion of the population in that group,

$$\mathrm{P}\left\{Y|do(X=1)\right\} = \sum_{z \in Z} \mathrm{P}\left\{Y|X=1,Z\right\} \mathrm{P}\left\{Z\right\} \tag{2.4}$$

We have seen from these two examples that the expression to estimate the causal effect of an intervention depends on the structure of the causal graph. There is a very powerful and general set of rules that specify how we can transform observational distributions into intervetional ones for a given graph structure. These rules are referred to as the Do-calculus [38]. We discuss them further in section **??**.

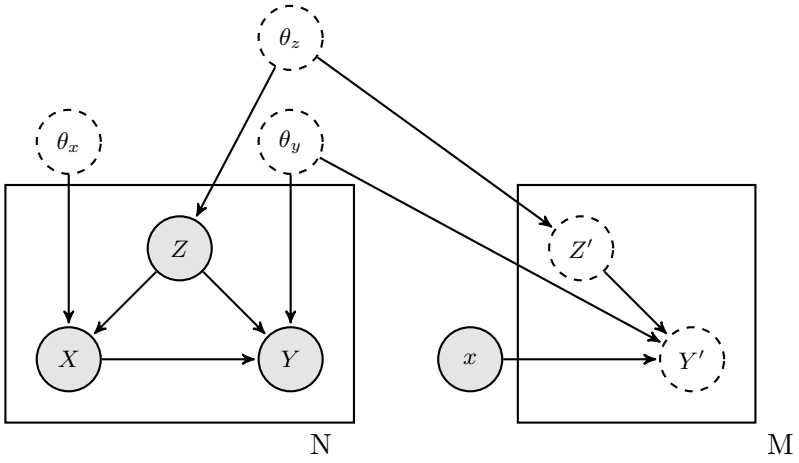
Formalising the definition of an intervention within the framework of causal graphical models provides us with an explicit mechanism to map information from one data generating process, the system pre-intervention, to another, the system post-intervention. The power of defining an intervention in this way stems from the number of things that are invariant between the two processes. All the (conditional) distributions for variables in the graph that were not directly set by the intervention are assumed not be changed by it.

We could represent problems of the type where we try to infer properties of the post-interventional system based on data generated by the pre-interventional distribution by explicitly representing both systems and what they have in common, see figure 2.5. This does not require any special framework or notation. The graphs in figure 2.5 are ordinary Bayesian networks. However, without a causal framework, we have to make assumptions about what will be invarient to the intervention specifically for each such problem we encounter. For complex problems, it is very difficult to conceptualise the assumptions we expect to hold without the benefit of a causal framework.

A causal bayesian network represents much more information than a bayesian network with identical structure. A causal network encodes all possible interventions that could be specified with the do-notation. For example, if the network in figure 2.4 were an ordinary bayesian network and all the variables were binary, the associated distribution could be described by 7 parameters. The equivalent causal bayesian network additionally represents the post-interventional distributions for six possible single variable interventions and twelve possible two variable interventions. Encoding all this information without the assumptions implicit in the causal bayesian network would require an additional 30 parameters. [3]

[3] After each single variable intervention we have a distribution over two variables, which can be represented by three parameters. After each two variable intervention, we have a distribution over one variables which requires one parameter. This takes us to a total of $6 * 3 + 12 * 1 = 30$ additional parameters.

Figure 2.5: Causal inference with ordinary bayesian networks. The plate on the left represents the observed data generated prior to the intervention and the plate on the right the data we anticipate obtaining after an intervention that the pre-interventional variable X to x. The assumptions characterised by this plate model correspond to those implied by the causal bayesian network in figure 2.4 for the intervention $do(X = x)$. As the networks in this figure are ordinary Bayesian networks, we could have represented the same information with a different ordering of the links within each plate. However, we would then have a complex transformation relating the parameters between the two plates rather than a simple invariance.

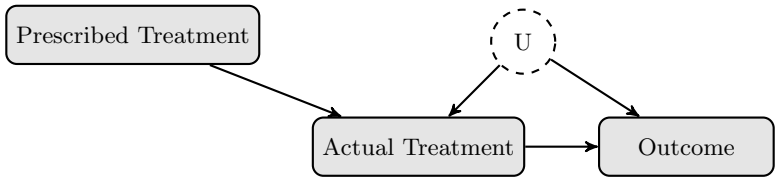


Causal bayesian networks are bayesian networks, so results that apply to bayesian networks carry directly across; the local Markov property states that variables are independent of their non-effects given their direct causes. Similarly the global Markov property and d-separation also hold in causal networks.

Limitations of causal bayesian networks

A number of criticisms have been levelled at this approach to modelling causality. One is that the definition of an intervention only in terms of setting the value of one or more variables is too precise and that any real world intervention will effect many variables in complex and non-deterministic ways [**?** 8]. However, by augmenting the causal graph with additional variables that model how interventions may take effect, the deterministic do operator can model more complex interventions. For example, in the drug treatment case, we assumed that all subjects complied, taking the treatment or placebo as assigned by the experimenter. But what if some people failed to take the prescribed treatment. We can model this within the framework of deterministic interventions by adding a node representing what they were prescribed (the intervention) which probabilistically influences the treatment they actually receive ,see figure 2.6. Note that the fact that we no longer directly assign the treatment opens the possibility that an unobserved latent variable could affect both the actual treatment taken and the outcome.

Figure 2.6: Randomised treatment with imperfect compliance



Another key issue with causal Bayesian networks is that they cannot handle cyclic dependen-

cies between variables. Such feedback loops are common in real-life systems, for example the relationship between supply and demand in economics or predator and prey in ecology. We might regard the underlying causal mechanisms in these examples to be acyclic; the number of predators at one time influences the number of prey in the next period and so on. However, if our measurements of these variables must be aggregated over timeframes that are longer than the scale at which these interactions occur the result is a cyclical dependency. Even were we able to measure on shorter timescales, we might then not have sufficient data on each variable for inference. Such problems have mostly been studied within the dynamic systems literature, typically focusing on understanding the stationary or equilibrium state of the system and making very specific assumptions about functional form in order to make problems tractable. [41] compare the equilibrium approach to reasoning about cyclic problems with strucural equation models, which we discuss in section 2.3 and which can be seen as bayesian causal networks with additional functional assumptions.

2.2 Counterfactuals

The Neyman-Rubin model [47, 48, 46, 49, 50] defines causality in terms of potential outcomes, or counterfactuals. Counterfactuals are statements about imagined or alternate realities, are prevalent in everyday language and may play a role in the development of causal reasoning in humans [63]. Causal effects are differences in counterfactual variables; what is the difference between what would happen if we did one thing versus what would happen if we did something else.

In example 1, the causal effect of the drug relative to placebo for person i is the difference between what would happen if they were given the drug, denoted ${y_i}^1$ versus what would happen if they got the placebo, ${y_i}^0$. The fundamental problem of causal inference is that we can only observe one of these two outcomes, since a given person can only be treated or not treated. The problem can be resolved if, instead of people, there are units that can be assumed to be identical or that will revert exactly to their initial state some time after treatment. This type of assumption often holds to a good approximation in the natural sciences and explains why researchers in these fields are less concerned with causal theory.

Putting aside any estimates of individual causal effects, it is possible to learn something about the distributions under treatment or placebo. Let Y^1 be a random variable representing the potential outcome if treated. The distribution of Y^1 is the distribution of Y if everyone was treated. Similarly Y^0 represents the potential outcome for the placebo. The difference between the probability of recovery, across the population, if everyone was treated and the probability of recovery given placebo is $\mathrm{P}\left\{Y^1\right\} - \mathrm{P}\left\{Y^0\right\}$. We can estimate, (from an experimental or observational study);

$$\mathrm{P}\left\{Y|X=1\right\}, \text{ the probability that those who took the treatment will recover}$$
$$\mathrm{P}\left\{Y|X=0\right\}, \text{ the probability that those who were not treated will recover}$$

Now, for those who took the treatment, the outcome *had* they taken the treatment Y^1 is the same as the observed outcome. For those who did not take the treatment, the observed outcome is the sames as the outcome *had* they not taken the treatment. Equivalently stated:

$$\mathrm{P}\left\{Y^0|X=0\right\} = \mathrm{P}\left\{Y|X=0\right\}$$
$$\mathrm{P}\left\{Y^1|X=1\right\} = \mathrm{P}\left\{Y|X=1\right\}$$

If we assume $X \perp\!\!\!\perp Y^0$ and $X \perp\!\!\!\perp Y^1$:

$$\begin{aligned} \mathrm{P}\left\{Y^1\right\} &= \mathrm{P}\left\{Y^1|X=1\right\} = \mathrm{P}\left\{Y|X=1\right\} \\ \mathrm{P}\left\{Y^0\right\} &= \mathrm{P}\left\{Y^0|X=0\right\} = \mathrm{P}\left\{Y|X=0\right\} \end{aligned}$$

$$\implies \mathrm{P}\left\{Y^1\right\} - \mathrm{P}\left\{Y^0\right\} = \mathrm{P}\left\{Y|X=1\right\} - \mathrm{P}\left\{Y|X=0\right\}$$

The assumptions $X \perp\!\!\!\perp Y^1$ and $X \perp\!\!\!\perp Y^0$ are referred to as ignore-ability assumptions [46]. They state that the treatment a each person receives is independent of whether they would recover if treated and if they would recover if not treated. This is justified in example 1 due to the randomisation of treatment assignment. In general the treatment assignment will not be independent of the potential outcomes. In example 2, the children who attended preschool may be more likely to have graduated highschool had they in fact not attended than the children who actually did not attend, $X \not\perp\!\!\!\perp Y^0$. Similarly, had the poorer children who did not attend pre-school attended they might not have done as well as the children who did in fact attend, $X \not\perp\!\!\!\perp Y^1$. A more general form of the ignore-ability assumption is to identify a set of variables Z such that $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$.

Theorem 5 (Ignore-ability)**.** *If $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$,*

$$\mathrm{P}\left\{Y^1\right\} = \sum_{z \in Z} \mathrm{P}\left\{Y|X=1,Z\right\} \mathrm{P}\left\{Z\right\} \tag{2.5}$$

$$\mathrm{P}\left\{Y^0\right\} = \sum_{z \in Z} \mathrm{P}\left\{Y|X=0,Z\right\} \mathrm{P}\left\{Z\right\} \tag{2.6}$$

Assuming that within each socio-economic status level, attendence at pre-school is independent of the likelihood of graduating high-school had a person attended, then the average rate of high-school graduation given a universal pre-school program can be computed from equation 2.5. Note, that this agrees with the weighted adjustment formula in equation 2.4.

Another assumption introduced within the Neyman-Rubin causal framework is the Stable Unit Treatment Value Assumption (SUTVA) [48]. This is the assumption that the potential outcome for one individual (or unit) does not depend on the treatment assigned to another individual. As an example of a SUTVA violation, suppose disadvantaged four year olds were randomly assigned to attend pre-school. The later school results of children in the control group, who did not attend, could be boosted by the improved behaviour of those did and who now share the classroom with them. SUTVA violations would manifest as a form of model misspecification in causal Bayesian networks.

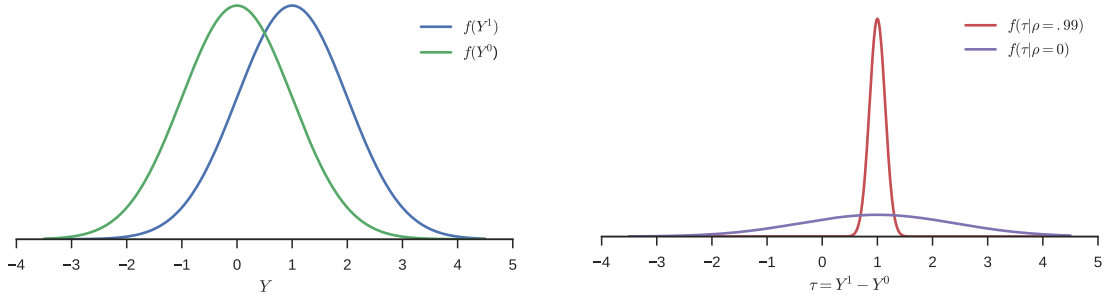
There are complex philosophical objections to counterfactuals arising from the way they describe alternate universes that were never realised. This makes it quite easy to (accidentally) make statements about counterfactuals that cannot be tested with empirical data. Consider the following example based on Dawid [12]. Again we have a drug, where the outcome for an individual if treated is represented by the counterfactual variable Y^1 and the outcome if not treated is Y^0. Suppose these counterfactual variables Y^1 and Y^0 are jointly normal with equal variance (for simplicity).

$$P(Y^1, Y^0) \sim N(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2\end{bmatrix}) \tag{2.7}$$

Their difference is also normal. Let $\tau = Y^1 - Y^0$,

$$P(\tau) = N(\mu_1 - \mu_0, 2\sigma^2(1-\rho)) \tag{2.8}$$

Figure 2.7: The distribution of individual treatment effects is not identifiable, even from a randomized controlled trial.



(a) Marginal distributions over the potential outcomes Y^1 and Y^0 for $\mu_1 = 1$, $\mu_0 = 0$ and $\sigma = 1$. The blue curve shows the distribution of Y if everyone were to be treated and the blue curve the distribution if no-one was treated.

(b) Two very different distributions of individual causal effects consistent with the potential outcome distributions.

From a (large) randomized controlled trial we can estimate the marginal distributions over the counterfactual variables, see figure 2.7a. These represent the distributions we would expect over the outcome Y if everyone were treated or not treated retrospectively. However, the distribution over the individual causal effects depends on ρ, see figure 2.7b. The key problem is that we can never observe the joint distribution over Y^1 and Y^0. As a result, ρ and thus the variance of τ is not identifiable, even from experimental data. **?**] argues that we should avoid using counterfactuals as they are defined in terms of (metaphysical) individual causal effects. He further points out that the interventional distributions in figure 2.7a, along with a loss function, contain all the information required to decide how to treat a new patient.

This result is unintuitive. It seems on the face of it that the distribution of individual causal effects is relevant to our decision making. If $\rho = 1$ then almost everyone benefits slightly from the treatment whilst if $\rho = 0$, there is a wide range, with some people benefiting a lot and others suffering significant harm. This confusion can be resolved by thinking about personalised rather than individual causal effects. It is entirely possible that potentially observable characteristics (such as gender, age, genetics, etc) affect how people will respond to the treatment. We can partition people into sub-populations on the basis of these characteristics and measure different *personalised* causal effects for each group. The variance of the potential outcome distributions $f(Y^1)$ and $f(Y^0)$ provides bounds on how much can be gained from further personalisation. The metaphysical nature of individual causal effects only arises when we are at the point where the only remaining variation is due to inherent randomness (or variables that we could not even in principle measure).

One way of looking at counterfactuals is as a natural language short hand for describing highly specific interventions like those denoted by the do-notation. Rather than taking about the

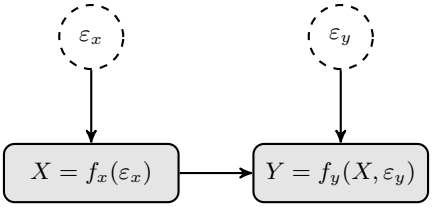
distribution of Y given we intervene to set $X = x$ and hold everything else about the system constant we just say what would the distribution of Y be had X been x. This is certainly convenient, if rather imprecise. However, the ease with which we can make statements with counterfactuals that cannot be tested with empirical data warrants careful attention. We should always be clear what assumptions we are making and have in mind if it is possible (at least in theory) to validate those assumptions.

2.3 Structural Equation models

Structural equation models (SEMs) describe a deterministic world, where some underlying mechanism or function determines the output of any process for a given input. The mechanism (but not the output) is assumed to be independent of what is fed into it. Uncertainties are not inherent but arise from unmeasured variables. Linear structural equation models have a long history for causal estimation [64, 20]. More recently, they have been formalised, generalised to the non-linear setting and connected to developments in graphical models to provide a powerful causal framework [38].

Mathematically, each variable is a deterministic function of its direct causes and a noise term that captures unmeasured variables. The noise terms are required to be mutually independent. If there is the possibility that an unmeasured variable influences more than one variable of interest in a study, it must be modelled explicitly as a latent (unobserved) variable. Structural equation models can be represented visually as a network. Each variable is a node and arrows are drawn from causes to their effects. Figure 2.8 illustrates the SEM for example 1.

Figure 2.8: SEM for example 1



This model encodes the assumption that the outcome y_i for an individual i is caused solely by the treatment x_i they receive and other factors ε_{y_i} that are independent of X. This is justifiable on the grounds that X is random. The outcome of a coin flip for each patient should not be related to any of their characteristics (hidden or otherwise). Note that the causal graph in figure **??** is identical to that the bayesian network for the same problem, figure 2.3. The latent variables ε_x and ε_y are not explicitly drawn in figure 2.3 as are captured by the probabilistic nature of the nodes in a Bayesian network, however adding them would not change the model.

Taking the *action* $X = 1$ corresponds to replacing the equation $X = f_x(\varepsilon_x)$ with $X = 1$. The function f_y and distribution over ε_y does not change. This results in the interventional distribution [4],

$$\mathrm{P}\left\{Y = y | do(X = 1)\right\} = \sum_{\varepsilon_y} \mathrm{P}\left\{\varepsilon_y\right\} \mathbb{1}\{f_y(1, \varepsilon_y) = y\} \tag{2.9}$$

The observational distribution of Y given X is,

[4] We have assumed the variables are discrete only for notational convinience

$$\mathrm{P}\{Y=y|X=1\} = \sum_{\varepsilon_x}\sum_{\varepsilon_y} \mathrm{P}\{\varepsilon_x|X=1\}\,\mathrm{P}\{\varepsilon_y|\varepsilon_x\}\,\mathbb{1}\{f_y(1,\varepsilon_y)=y\} \tag{2.10}$$

$$= \sum_{\varepsilon_y} \mathrm{P}\{\varepsilon_y\}\,\mathbb{1}\{f_y(1,\varepsilon_y)=y\}\,, \text{ as } \varepsilon_x \perp\!\!\!\perp \varepsilon_y \tag{2.11}$$

The interventional distribution is the same as the observational one. The same argument applies to the intervention $do(X=0)$ and so the causal effect is just the difference in observed outcomes as found via the causal Bayesian network and counterfactual approaches.

The SEM for example 2 is shown in figure 2.9. Intervening to send all children to pre-school replaces the equation $X = f_x(Z, \varepsilon_x)$ with $X = 1$, leaving all the other functions and distributions in the model unchanged.

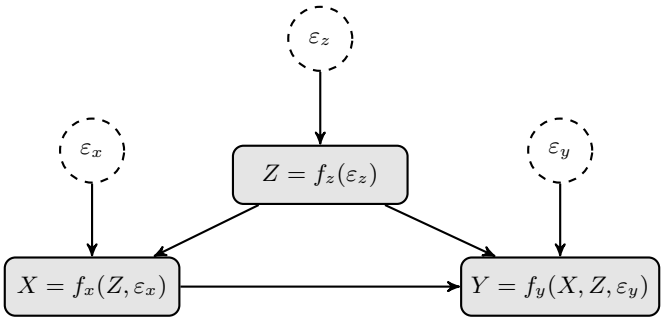
$$\mathrm{P}\{Y=y|do(X=1)\} = \sum_{z}\sum_{\varepsilon_y} \mathrm{P}\{z\}\,\mathrm{P}\{\varepsilon_y\}\,\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\} \tag{2.12}$$

$$= \sum_{z} \mathrm{P}\{z\} \underbrace{\sum_{\varepsilon_y} \mathrm{P}\{\varepsilon_y\}\,\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\}}_{\mathrm{P}\{Y=y|X=1,Z=z\}} \tag{2.13}$$

Equation 2.13 corresponds to equations 2.4 and 2.5. It is not equivalent to the observational distribution, given by;

$$\mathrm{P}\{Y=y|X=1\} = \sum_{z}\sum_{\varepsilon_y} \mathrm{P}\{z|X=1\}\,\mathrm{P}\{\varepsilon_y\}\,\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\} \tag{2.14}$$

Figure 2.9: SEM for example 2



Structural equation models are generally applied with strong constraints on the functional form of the relationship between the variables and noise is typically assumed to be additive, $X_i = f_i(\cdot) + \varepsilon_i$. A structural equation model with N variables resembels a set of N simultaneous equations, with each variable playing the role of the dependent (left hand side) variable in one equation. However a SEM is, by definition, more than a set of simultaneous equations. By declaring it to be structural we are saying that it represents assumptions about the relationships between variables. When we visualise the model as a network the absence of an arrow between two variables encodes the assumption that one does not cause the other. The similarity between the notation used to describe and analyse structural equation models and simultaneous equations, combined with a reluctance to make explicit statements about causality has led to some confusion in the interpretation of SEMs [23, 38].

2.4 Comparing and unifying the models

Remarkably for models developed relatively independently in fields with very different approaches and problems, the models we have discussed can be nicely unified for interventional queries (those that can be expressed with the do-notation). If the network for a structural equation model is acyclic, that is if starting from any node and following edges in the direction of the arrows you cannot return to the starting point, then it implies a recursive factorisation of the joint distribution over its variables. In other words, the network is a causal Bayesian network. All of the results that apply to causal Bayesian networks also apply to acyclic structural equation models. Taking an action that sets a variable to a specific value equates to replacing the equation for that variable with a constant. This corresponds to dropping a term in the factorisation and the truncated product formula (equation 2.3). Thus, the interventional query $P(Y|do(X))$ is identical in these two frameworks. We can also connect this to counterfactuals via:

$$\begin{aligned} Y^0 &\equiv P(Y|do(X=0)) \\ Y^1 &\equiv P(Y|do(X=1)) \end{aligned} \tag{2.15}$$

The assumption $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$, stated for our structural equation model, translates to $X \perp\!\!\!\perp (Y^0, Y^1)$ in the language of counterfactuals. When discussing the counterfactual model, we actually made the slightly weaker assumption:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1 \tag{2.16}$$

It is possible to relax the independence of errors assumption for SEMs to correspond exactly with the form of equation (2.16) without losing any of the power provided by d-separation and graphical identification rules [45]. The correspondence between the models for interventional queries (those that can be phrased using the do-notation) makes it straightforward to combine key results and algorithms developed within any of these frameworks. For example, you can draw a causal graphical network to determine if a problem is identifiable and which variables should be adjusted for to obtain an unbiased causal estimate. Then use propensity scores[46] to estimate the effect. If non-parametric assumptions are insufficient for identification or lead to overly large uncertainties, you can specify additional assumptions by phrasing your model in terms of structural equations. The frameworks do differ when it comes to causal queries that involve joint or nested counterfactuals and cannot be expressed with the do-notation. These types of queries arise in the study of mediation [39, 28, 60] and legal discrimination [38].

In practice, differences in focus and approach between the fields in which each model dominates eclipse the actual differences in the frameworks. The work on causal graphical models [38**?**] focuses on asymptotic, non-parametric estimation and rigorous theoretical foundations. The Neyman-Rubin framework builds on the understanding of randomised experiment and generalises to quasi-experimental and observational settings, with a particular focus on non-random assignment to treatment. This research emphasises estimation of average causal effects and provides practical methods for estimation, in particular, propensity scores; a method to control for multiple variables in high dimensional settings with finite data [46]. In economics, inferring causal effects from non-experimental data to support policy decisions is central to the field. Economists are often interested in broader measures of the distribution of causal effects than the mean and make extensive use of structural equation models, generally with strong parametric assumptions [22]. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models.

2.5 What does a causal model give us? Resolving Simpson's paradox

We will now demonstrate our new notation and frameworks for causal inference to resolve a fascinating paradox, noted by Yule [65], demonstrated in real data by Cohen and Nagel [10] and popularised by Simpson [55]. The following example is adapted from Pearl [38]. Suppose a doctor has two treatments, A and B, which she offers to patients to prevent heart disease. She keeps track of which medication her patients choose and whether or not the treatment is successful. She obtains the results in table 2.1.

Table 2.1: Treatment results

Treatment	Success	Fail	Total	Success Rate
A	87	13	100	87%
B	75	25	100	75%

Drug A appears to perform better. However, having read the latest literature on how medications affect men and women differently, she decides to break down her results by gender to see how well the drugs perform for each group and obtains the data in table 2.2.

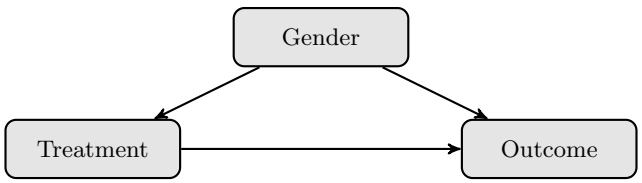
Table 2.2: Treatment results by gender

Gender	Treatment	Success	Fail	Total	Success Rate
M	A	12	8	20	60%
M	B	56	24	80	70%
F	A	75	5	80	94%
F	B	19	1	20	95%

Once the data is broken down by gender, Treatment B looks better for both men *and* women. Suppose the doctor must choose only one drug to prescribe to all her patients in future (perhaps she must recommend which to subsidise under a national health scheme). Should she choose A or B? The ambiguity in this question lies at the heart of Simpson's paradox. How does causal modelling resolve the paradox? The key is that the doctor is trying to choose between *interventions*. She wants to know what the success rate will be if she changes her practice to give all the patients one drug, rather than allowing them to choose as currently occurs.

Let's represent the treatment by the variable T, the gender of the patient by Z and whether or not the treatment was successful by Y. The doctor cares about $\mathrm{P}\{Y|do(T)\}$, not the standard conditional distributions $\mathrm{P}\{Y|T\}$. Unfortunately, the data in tables 2.1 and 2.2 is insufficient to enable estimation of the interventional distribution $\mathrm{P}\{Y|do(T)\}$ or determine if $do(T = A)$ is better or worse than $do(T = B)$. Some assumptions about the causal relationships between the variables are required. In this example, it seems reasonable to conclude that gender may affect the treatment chosen and the outcome. Assuming there are no other such confounding variables (for example income) then we obtain the causal network in figure 2.10.

Figure 2.10: An example of Simpson's Paradox

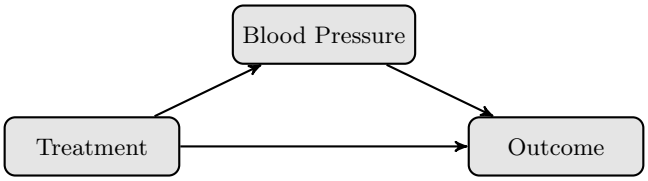


With this model, women are more likely to choose treatment A and are also more likely to recover than men regardless of the treatment they receive. Knowing a patient took drug A indicates they are more likely to be female. When we compare the group of people who took A against those who took B, the effect of the higher share of females in the first group conceals the greater benefit of drug B leading to an apparent reversal in effectiveness. However, when the doctor intervenes to set the treatment each person recieves there will no longer be a link from gender to treatment. So in this case she should choose which drug to prescribe from the gender specific table (and weight by the proportion of the population that belongs to each gender). Drug B is the better choice.

$$\mathrm{P}\left\{Y|do(T)\right\} = \mathrm{P}\left\{Y|T, female\right\}\mathrm{P}\left\{female\right\} + \mathrm{P}\left\{Y|T, male\right\}\mathrm{P}\left\{male\right\} \tag{2.17}$$

Is the solution to Simson's paradox to always to break down the data by as many variables as possible? No. Suppose we have the identical data as in 2.1 and 2.2 but replace the column name 'gender' with 'blood preassure', 'M' with 'high' and 'F' with 'normal'. This is a drug designed to prevent heart disease. One pathway to doing so might well be to lower blood pressure. Figure 2.11 shows a plausible causal graph for this setting. It differs from the graph in figure 2.10 only in the direction of a single link. Now, however table 2.2 tells us that people who took treatment A had better blood pressure control and better overall outcomes. In this setting $\mathrm{P}\left\{Y|do(T)\right\} = \mathrm{P}\left\{Y|T\right\}$. Drug A is the better choice.

Figure 2.11: An example of Simpson's Paradox



Note that we have not changed the data itself, only the description of the variables that it is associated with. This illustrates that the resolution to Simpson's paradox lies fundamentally not in the data, but in the assumptions we are willing to make. From a purely statistical viewpoint there is no paradox. The reversal just stems from the mathematical property of ratios expressed in equation 2.18 and represented graphically in figure 2.12. The paradox only arises when we attempt to use the data to select an intervention and is resolved when we apply a causal approach to do so.

$$\exists\left\{N_1, ...N_4, N_1'...N_4'\right\} \in \mathbb{N}: \ \frac{N_1}{N_2} < \frac{N_1'}{N_2'}, \ \ \frac{N_3}{N_4} < \frac{N_3'}{N_4'} \text{ and } \frac{N_1 + N_3}{N_2 + N_4} > \frac{N_1' + N_3'}{N_2' + N_4'} \tag{2.18}$$

There are many other plausible causal graphs for both scenarios above. Perhaps income affects drug choice as well as gender, or gender might affect treatment choice and blood preassure control given treatment, etc. Causal modeling provides a powerful tool to specify such assumptions and to determine how to estimate causal effects for a given model see section 3.1.

Figure 2.12: Simpson's reversal visualized. The ratios involving N_i' are steeper than than those involving N_i for both the blue and green vectors. However, when we sum them, the ratio is steeper for the un-primed variables.