Causal Inference using Directed Acyclic Graphical (DAG) Models

A report presented as a part of the course
Categorical Data Analysis, Fall 2019

Subhrajyoty Roy, MB1911
Arindam Roy Chowdhury, MB1919
Abhinandan Dalal, MB1920

Instructor: Soumendu Sundar Mukherjee
Programme : Master of Statistics (M.Stat. Year I)
Institution: Indian Statistical Institute, Kolkata
Date: December 20, 2019

Contents

1 Introduction

The questions that motivate most studies in the health, social and behavioral sciences are not associational but causal in nature. From policy formulation in social sciences to assessing effectiveness of newly created drug in biological science, causal inference is very much demanded in different disciplines of science.

In the usual statistical techniques, we proceed by following structures;

1. First, we specify a model to describe the data.
2. If possible, we try to see whether the model is valid for the observed data that we have at hand, using outlier detection or through means of hypothesis testing. If the model is not adequate, revise the model.
3. Then we estimate various parameters of the specification for the model, through means of the observed data.
4. Finally, we use the estimated parameters to fully specify the model, and use that to answer the required question which we are trying to answer, i.e. to meet the requirements specified by the objective of the study.

However, such a statistical theory fails to answer the question mentioned before, due to the assumption that it only captures the essence of the behaviour of samples following a distribution governed by the model specification. There is nothing in a distribution function (as used in classical theory of Statistics) to tell us how that distribution would differ if external conditions were to change, say from observational to experimental setup, because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions which identify relationships that remain invariant when external conditions change. Causal techniques are able to make inference about the sample even under the situation when underlying distribution changes due to the effect of treatments or interventions on the sampling units.

2 Simpson's Paradox in the Light of Causality

Simpson's paradox is a popular type of paradox which appears in Statistics, where a trend or some pattern emerges within each individual groups, but reverses when all groups are combined for making inference. To show how Simpson's paradox can be laid out in the light of causal inference, we consider two examples.

Example 1. *Consider a dataset with variables Age, Cholesterol and Exercise. It is known that within a particular age group, Exercise is negatively associated with Cholesterol. Whereas, if we compare between two age groups, then the younger people tend to have lower cholesterol level than the older people. Therefore, if we look at the scatter diagram of Age and Cholesterol, we would have a diagram as shown in Figure 1. So, the particular question that we are trying to answer is whether a doctor should prescribe a patient to do exercise to reduce cholesterol on the basis of Figure 1.*

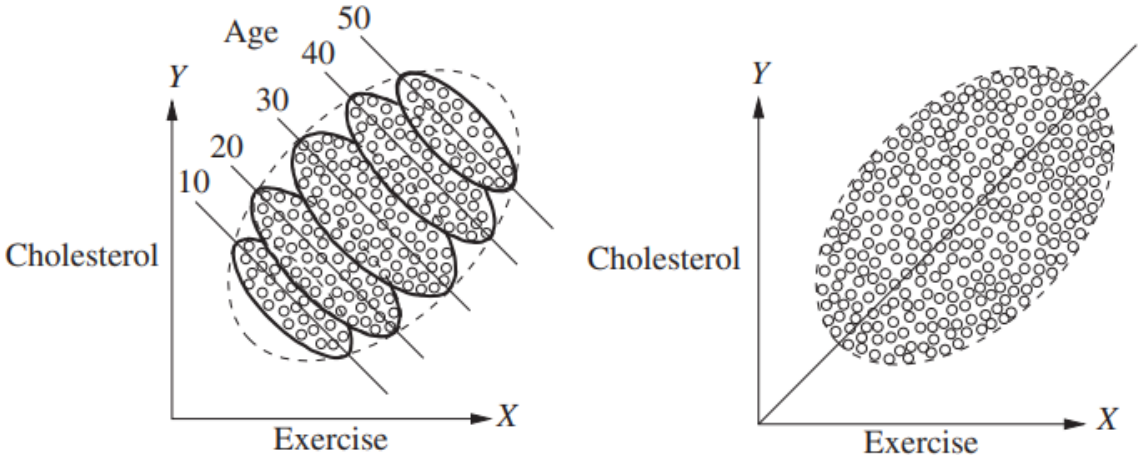


Figure 1: Scatter diagram for showing relationship between Age, Cholesterol and Exercise: An example of Simpson's Paradox

Example 2. *Let us say there is a new drug developed as a possible cure for a disease. However, the drug has an indirect effect, which heavily reduces the blood pressure, and suppose that reduction of blood pressure helps to cure the disease to a great extent. However, there is also a slight negative side effect of the drug which worsens the disease. Hence, if we consider a scatter diagram between the dosage of the drug in x-axis and some measure of cureness of the disease in y-axis, then we would obtain a diagram similar to Figure 1, where within each fixed level of blood pressure, the effect of the drug is negative. Here, also the question we wish to answer is whether a doctor should prescribe the drug to a patient for curing the disease.*

So, both the Example 1 and Example 2 results in a diagram similar to Figure 1. However, we know that the answers to the questions in the examples are very different. In Example 1, we should condition on the variable Age group, which would enable the doctor to prescribe exercising as a way to reduce cholesterol level. On the other hand, in Example 2, we should not condition on the variable blood pressure, since the negative side effect is smaller in magnitude than the indirect positive effect through blood pressure. Therefore, a doctor should prescribe the drug as a cure to the disease without looking at the individual groups, but rather looking at the whole picture at once.

The rationale behind choosing such different methods to answer the question, is not at all clear from point of view of classical statistics. However, once we know the causality structure of the data, it would be seen that in Example 1, the variable age group was a confounder, while in Example 2, the variable blood pressure is a mediating variable. Why conditioning on confounder makes sense, and why conditioning on a mediating variable makes sense of the true underlying structure of the data, these questions can be answered through the mathematical framework of Causal inference, which will be discussed in subsequent sections.

3 Causal Assumptions

Before going into devising a methodology for identifying causal effects, we first need to be clear on what we are aiming to do, and what we treat as given. This is devised in the following Causal Assumptions:

- **Stable Unit Treatment Value Assumption** (SUTVA)
  - Units do not interfere with each other.
  - Treatment assignment of one unit does not affect outcome of another (even no spillovers).
  - Only one version of treatment (else difficult to associate outcome to treatment).
- **Consistency**: Potential outcome under treatment $A = a$ is equal to observed outcome if treatment received is indeed a.
- All relevant variables are included in the model (possibly unobserved).

The skeptic reader may question that, how does the causality contribute anything new, if these assumptions are dealt into the fabric. The third assumption, in particular, can be the most troubling to the reader. It is true that most of these assumptions are, in general not testable. However, even if all the relevant variables are known, it is not enough to identify the underlying causal structures that seemed to generate the model. That is, whether A causes B or B causes A may not be known, even if A and B are included in the model. Our work shall be to obtain these causal structures, and obtain testable implications based on them.

4 Causal Inference and DAG Models

It should be pretty much clear that to perform Causal Inference, we need to have something more than the data itself. The reason is that, if we only have the data, then it is absolutely not possible to know how many other variables are there which affect the underlying distribution that generates this data, and it is also not possible to know about how this distribution will react to the changes in variables not included in the dataset. For this reason, we shall need another component in performing Causal inference, which is basically a **Structural Causal Model**.

Structural Causal Model is a Directed Acylic Graph (DAG) which is built as follows;

1. The vertex set V consists of all of the variables in the data, including all other variables which can be measured, or observed or can be inferred to influence any of the variables in the data, in direct or indirect manner.
2. The edge set E consists of all the directed edges (X, Y), if Y is a direct cause of variable X. To mean by direct cause, it is assumed that nature sees the value of X and then decides on the value of Y. In mathematical terms, we should have;
$$\mathbb{P}(Y = y \mid X = x) \neq \mathbb{P}(Y = y)$$
3. For each vertex X, there is an incoming edge to X from U_X which denotes the effect of all unexplained variables (or errors) that directly influences the variable X.
4. All the errors or unexplained variables are thought to be independently distributed.

This Directed Acylic Graph can be thought as a generating mechanism for the observed dataset. It is as if nature first generates the unexplained variables U_X's independently, and then move through the DAG to decide the values of the direct causes of those unexplained variables, and then direct causes of those, and the process continues until all the variables have been determined by the nature.

It is a natural question to ask why such a DAG model is true at all, and why such a generating mechanism is at all a possible action for nature to follow up with? To this end, we consider the following example.

Example 3. *Let us consider three variables under study,* ***`Smoking habit, Lung Cancer`*** *and* ***`Gender`****. Let, X denote the variable Smoking habit in a way that it takes value 1 if a person smokes 1 or more cigarettes per day. Let Y be a variable that takes value 0 or 1 based on whether a person has lung cancer of not. And let Z be the varible denoting gender of the person, with the interpretation that it takes value 0 for females and 1 for males.*

If we wish to draw a causal diagram for this type of setup, the diagram would possibly look like Figure 2.

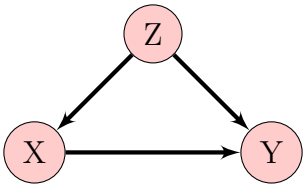


Figure 2: SCM for relationship between Smoking habit, Lung Cancer and Gender

Note that it is obvious that Smoking habit or having Lung Cancer does not cause the Gender, hence the arrow from X to Z or the arrow from Y to Z is not possible. However, there might be an arrow from X to Y, which is worth investigating, under the setup of the study.

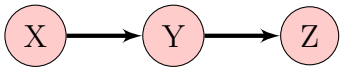
Coming back to the study of SCM, it is worthwhile to note that the underlying distribution associated with the data is assumed to be factorized over the directed acyclic graph. This factorization actually helps to connect the two components of the data, i.e. the DAG which builts the causality flow diagram of the variables presented in the data, and the distribution which generates the observed data. Since the factorization property is assumed to hold, hence all of the nice Markov properties are enjoyed by these Structural Causal models.

Let, $X_1, X_2, \dots X_n$ be the n variables presented in the DAG as vertices. This factorization property leads to the following decomposition of the joint distribution of these random variables;

$$\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \dots X_n = x_n\right) = \prod_i \mathbb{P}\left(X_i = x_i \mid Pa_i\right)$$

where Pa_i is the set of parent nodes for the variable X_i in the DAG.

Example 4. *Consider for example, we have a SCM which looks like Figure 4.*



Note that, if we consider the joint distribution of (X, Y, Z), then;

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y \mid X)\mathbb{P}(Z \mid X, Y)$$

However, according to factorization property, we would have;

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y \mid X)\mathbb{P}(Z \mid Y)$$

These two equations together would imply; $\mathbb{P}(Z \mid X, Y) = \mathbb{P}(Z \mid Y)$, which basically says that X and Z are conditionally independent given Y, which is same as the Pairwise Markov Property in this graph.

5 Special Types of Graphical Models

5.1 Chain

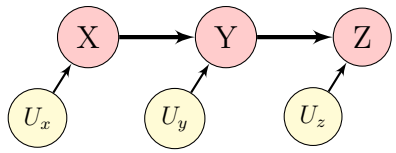


Figure 3: Chain path diagram

(i) $Y \not\perp\!\!\!\perp X$

(ii) $Z \not\perp\!\!\!\perp Y$

(iii) X and Z are likely to be dependent

(iv) $X \perp\!\!\!\perp Z \mid Y$

U_x, U_y and U_z are unexplained random errors

Example 5. *Consider a switch, connected to a bulb via a wire. Let X be a random variable which takes value 1 or 0 depending on whether the switch is on or off. Similarly, Y = indicator of current in circuit and Z = status of light bulb (on or off). Clearly, this is a chain $X \to Y \to Z$. The bulb, wire and switch may not always work perfectly. There may be dysfunctional, or the wire may get damaged. These add to the random errors U_i.*

If a variable Y depends on X, and Z depends on Y, then Z is *likely* to depend on X. But this is not always true. Consider the following example:

Example 6. *Let U_x, U_y, U_z be binary variables.*

$$X = U_x$$

$$Y = \begin{cases} a & \text{if } X = 1 \text{ and } U_y = 1 \\ b & \text{if } X = 2 \text{ and } U_y = 1 \\ c & \text{if } U_y = 2 \end{cases}$$

$$Z = \begin{cases} i & \text{if } Y = c \text{ or } U_z = 1 \\ j & \text{if } U_z = 2 \end{cases}$$

It is easy to see that the above is a chain where $X \perp\!\!\!\perp Z$.

5.2 Fork

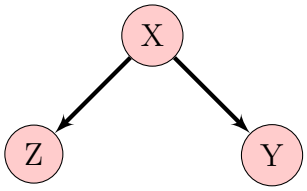


Figure 4: Fork path diagram

(i) $X \not\perp\!\!\!\perp Y$

(ii) $X \not\perp\!\!\!\perp Z$

(iii) Y and Z are likely to be dependent

(iv) $Y \perp\!\!\!\perp Z \mid X$

Example 7. *Let X denotes the weather temperature, Z denotes the number of ice cream sales in a region and let Y denotes the number of person who gets drowned when swimming. It is evident that if X is high, i.e. the weather is hot, then both the ice-cream sale and the number of person drowned when swimming both are high, as people would like to go more to swim in summer. However, these quantities are seemingly unrelated when the temperature is kept fixed at a certain level.*

Also, it is clear that both of these phenomenon occur due to the hot weather, but the temperature cannot be caused because of change in ice-cream sale or change in number of deaths by drowning. Hence, the resulting graph would possibly yield a fork as shown in Figure 4.

5.3 Colliders

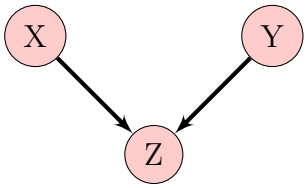


Figure 5: Collider path diagram

(i) $X \not\perp\!\!\!\perp Z$

(ii) $Y \not\perp\!\!\!\perp Z$

(iii) $X \perp\!\!\!\perp Y$

(iv) $X \not\perp\!\!\!\perp Y \mid Z$

Example 8. *Consider an University that gives scholarships to a student if either the student has good grades, or, the student has special talent in music. Now, take:*

$$\begin{aligned} Z = &\ \textit{Received scholarship or not} \\ Y = &\ \textit{Whether the student has musical talent} \\ X = &\ \textit{Grade of that student} \end{aligned}$$

In the above example, X causes Z, and Y also causes Z. Now, school grades are independent of musical talent(*assumption*). But, given a student received scholarship, $X \not\perp\!\!\!\perp Y$, because, consider a student with low grades receiving scholarship. Then, he must be musically talented.

5.4 d-separation

Generally, causal models are not as simple as the ones discussed above. It is rare for a graphical model to have a single path between variables. In most graphical models, pairs of variables will have multiple possible paths connecting them. The objective is to

have a criterion, that can be used to predict dependencies that are shared by all data sets generated by the graph. *d-separation* is one such criterion based on the structures discussed above (*d* stands for "directional"). Two nodes X and Y are *d-separated* if every path between them are *blocked*. If any one path between X and Y are *unblocked*, then X and Y are *d-connected*.

Definition 1. *A path p is blocked by a set of nodes Z if and only if:*

1. *p contains a chain of nodes* $A \to B \to C$ *or a fork* $A \leftarrow B \to C$ *such that the middle node* B *is in* Z*.*

   OR

2. *p contains a collider* $A \to B \leftarrow C$ *such that the collision node* B *is not in* Z *and no descendant of* B *is in* Z

If Z blocks every path between nodes X and Y, then X and Y are d-separated conditional on Z, and thus are independent conditional on Z.

Example 9.

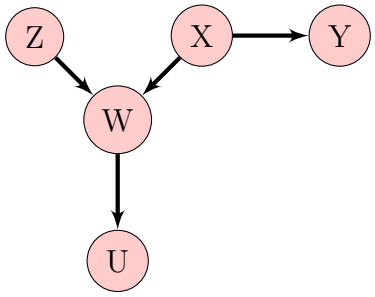


Figure 6: A graphical model containing a collider with a child and a fork

In the graph in Figure 6, we have the following relationships:

(i) Using an empty conditioning set, Z and Y are *d-separated*. This is because they have only one path between them and that is blocked by the collider $Z \to W \leftarrow X$.

(ii) Taking the conditioning set = $\{W\}$, we have Z and Y are *d-connected*, because there is a fork $(X) \notin \{W\}$, and the only collider W is in the set.

(iii) Now, considering $\{W, X\}$ as our conditioning set, Z and Y are *d-separated*, as the path between them is blocked by the first criterion. That is, there is now a fork $(X) \in \{X, W\}$, the conditioning set.

6 Model Testing

Suppose we have a graph G that we believe might have generated a data set S. The, *d-separation* will tell us which variables must be independent given which other variables. Now, conditional independence is something we can test for using a data set. Suppose we list the *d-separation* conditions in G, and note that variables A and B must be independent conditional on C. Then, suppose we estimate the probabilities based on S, and discover that the data suggests that A and B are not independent conditional on C. We can then reject G as a possible causal model for S.

Example 10. *Consider the following graph:*

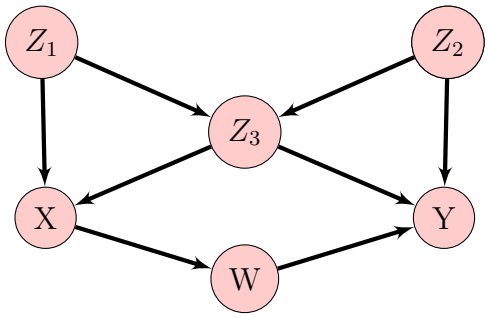


Figure 7: A Causal Graph to be tested

Among the many conditional independencies advertised by the model, we find that W and Z_1 are independent given X, because X *d-separates* W from Z_1. Now suppose we regress W on X and Z_1:

$$w = r_X x + r_1 z_1$$

and test for $H_0 : r_1 = 0$. If H_0 gets rejected, we know that W depends on Z_1 given X and, consequently, that the model is wrong. Not only we know that the model is wrong, but we also know where it is wrong; the true model must have a path between W and Z1 that is not *d-separated* by X.

If we have enough computational resources, based on a data set, we can test all possible conditional independencies. This would lead to rejection of many SCM graphs, and we would be left with a few graphs in an equivalence class.

However, this method has several cons. First of all, this method is inherently parametric, hence, if the parameter for any part of the model cannot be estimated reliably, the full procedure breaks down. However, we can instead do the estimation of the causal graph based on the d-separations, which provide several advantages.

- The d-separations, in principle, do not rely on parametric assumptions, based on just the graphs. Thus, even if the model specification for a local area goes wrong, it can be corrected locally.
- Moreover, d-separations tests locally, rather than globally. This allows us to identify local areas where the causal structure goes wrong, and allows to repair them, rather than starting afresh.
- Furthermore, even if the coefficients in one part of the model cannot be identified, for whatever reason, it yet allows us to infer about a different part of the graph.

A general strategy to find out the causal graph from the observed variables is to find out the underlying undirected graph from the marginal dependencies. This is commonly known as the *skeleton* of the causal graph. Then the possible causal graphs can be tested for using *d*-separations.

7 Estimating Effect of Intervention

7.1 Deriving Adjustment Formula

The particular reason why causal inference is studied extensively in different subjects like economics, biology, social sciences etc. is because we wish to see to how much of an extent one variable directly causes the other variables, under the assumption of *ceteris paribus*, i.e. under the assumption that all other variable does not change. However, due to the nature of causality, the change in one variable, would certainly have impact on all its descendants, hence we would end up violating this *ceteris paribus* assumption.

Therefore, in order to assess some inference based on causality, we need a convert the setup of the data to an experimental setup, where we would have a controlled group and a treatment group, all with same or similar characteristics for all other variables (as if we are simulating two different universes with same characteristics), but have a change in only one variable. Then, the effect on the response would be entirely attributed to the change of that variable, which would help us determining specific effects, as well as help us formulate policies with the help of the study. For example, in a drug testing setup, we are required to have two groups of patients, one group being treated with placebos and the other group being treated with the drug under study. In such a controlled experiment, we would be able to assess where the drug has any effect of curing the disease it is meant to cure. However, such an intervention on a variable by fixing its value with a controlled experiment may not be physically feasible.

For instance, consider the Example 7. Note that, it is physically impossible to perform an experiment where we fix the ice cream sale at 0, and then look at whether the number of person drowned when swimming is becoming small, hence try to assess the effect of ice cream sale on the number of drowned person. But performing such experiment would require shutting down all ice cream shops in the town, or a region or a country. However, intervention is a technique which actually imitates this experimental setup just by changing the causal graph, at a theoretical level rather than an experimental level.

If we intervene on a variable X i.e. we fix the value of variable X via an experiment, then all other variables that are a direct cause of X does not change, and hence the interpretation that those direct causes actually causes the value of X to change goes away. Hence we can simply study the effect of experimentation on X by removing all edges coming into the node corresponding to X in the associated DAG. Then we can simply act as an attorney for nature and choose the error (or unobservable) distribution to be of our choice, specifically to be degenerate at the value required.

When we intervene on X, we denote this by using **do** operator. Note that, $do(\cdot)$ operator is not same as conditional operator. For instance, $do(X = x)$ means that for each person in the population we have the value of variable X is set to x, while conditioning on the event $X = x$ means to look at the sub-population for which the value of the variable X turns out to be x.

Example 11. *Let X, Y, Z be three binary variables, representing whether a patient is given a specific drug or not, whether a patient has recovered from the disease under consideration and the gender of the patient respectively. In such a setup, the causal model can be thought to be same as a fork which Z as a mediator (i.e. the variable at the root of the fork), with an extra edge from X to Y, representing that the drug might have an effect on the recovery of the patient.*

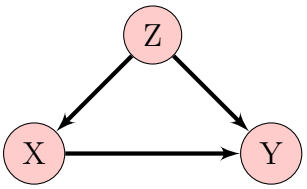


Figure 8: Full DAG for Drug testing setup

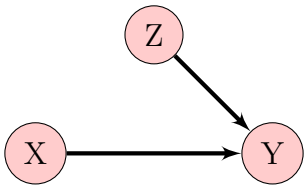


Figure 9: Modified DAG for Drug testing setup after intervention on X

The task is to estimate the causal effect of the drug on recovery, i.e.

$$\mathbb{P}(Y = 1|do(X = 1)) - \mathbb{P}(Y = 1|do(X = 0))$$

We focus on estimating the general term, $\mathbb{P}(Y = y|do(X = x))$. Now, when we intervene on X, i.e. use $do(X = x)$, we remove the incoming edge coming to X, and obtain a modified DAG as shown in Figure 9.

Now, note that, in both the full DAG and modified DAG, the marginal distribution of Z is same. Hence, denoting the probability measure in modified DAG by $\mathbb{P}_m$, we have;

$$\mathbb{P}_m(Z = z) = \mathbb{P}(Z = z)$$

Also note that, the arrows entering into Y remains same for both the full DAG and modified DAG. Hence,

$$\mathbb{P}_m(Y = y \mid X = x, Z = z) = \mathbb{P}(Y = y \mid X = x, Z = z)$$

Also, Z and X are d-seperated in the modified DAG, as the modified DAG essentially forms a collider. Hence, $\mathbb{P}_m(Z = z \mid X = x) = \mathbb{P}_m(Z = z)$.

Therefore, the intervention effect can be expressed as;

$$\begin{aligned}
\mathbb{P}\left(Y = y \mid do(X = x)\right) &= \mathbb{P}_m\left(Y = y \mid X = x\right) \quad \text{by definition of do operator} \\
&= \sum_z \mathbb{P}_m\left(Y = y \mid Z = z, X = x\right) \mathbb{P}_m\left(Z = z \mid X = x\right) \\
&= \sum_z \mathbb{P}\left(Y = y \mid Z = z, X = x\right) \mathbb{P}_m\left(Z = z\right) \\
&= \sum_z \mathbb{P}\left(Y = y \mid Z = z, X = x\right) \mathbb{P}\left(Z = z\right) \qquad (1)
\end{aligned}$$

Note that, the last term is a formula with all quantities of the full DAG model, hence the intervention effect can be estimated from the data alone. Note that, the above term is essentially a weighted sum of conditional probabilities, with weights being the probabilities of a conditioning variable. Hence, the conditioning variable works as if the effect of do operator is obtained through adjusting for other variable which affects both X and Y. Hence, the above formula is called **Adjustment Formula**.

Theorem 1 (The Adjustment Formula)**.** *Given a graph G, let X be a set of variables on which we are performing intervention upon, and let PA denote the set of variables which are parents of some variable in X, then the intervention effect (or causal effect) is given by;*

$$\mathbb{P}\left(Y = y \mid do(X = x)\right) = \sum_{z} \mathbb{P}\left(Y = y \mid X = x, PA = z\right) \mathbb{P}\left(PA = z\right)$$

Proof. We shall use the decomposition of the probability measure (i.e. the factorization property) to proof the theorem.

Note that, for the full model, we have the decomposition;

$$\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n\right) = \prod_{i} \mathbb{P}\left(X_i = x_i \mid Pa_i\right)$$

where the symbols are defined as before. When we intervene on a set of variables X, then there are 3 types of effects which may potentially affect the original DAG.

1. Edges for which none of the endpoint is an vertex belong to the set X changes from full DAG to modified DAG.
2. Edges for which we have a vertex of X at the parent edge, they remain same in full DAG and modified DAG.
3. Edges for which we have a vertex of X at the child node disappears in the modified DAG.

Therefore, in the modified model, we would simply have a product decomposition like;

$$\mathbb{P}_m\left(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n \mid do(X = x)\right) = \prod_{i:X_i \notin X} \mathbb{P}\left(X_i = x_i \mid Pa_i\right) \prod_{i:X_i \in X} \mathbb{P}(X_i)$$

i.e. the product extends over only those variables for which $X_i \notin X$, pertaining to the effect in situation 3, and for the variable in X, it just obtains the marginal distribution. Now, letting $V(G) = Y \cup X \cup PA \cup W$, we have;

$$\begin{aligned}
&\mathbb{P}_m\left(Y = y, X = x, PA = z, W = w\right) \\
=&\mathbb{P}(X = x)\mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z \mid PA_{PA}) \ldots \\
=&\mathbb{P}(X = x)\mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z, W = w) \\
=&\mathbb{P}(X = x)\mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z)\mathbb{P}(W = w \mid PA)
\end{aligned}$$

Note that, the term $\mathbb{P}(X = x \mid PA)$ does not appear in the above factorization as we apply do operator on X. Then,

$$\begin{aligned}
\mathbb{P}_m(PA = z) &= \sum_{x,y,w} \mathbb{P}_m\left(Y = y, X = x, PA = z, W = w\right) \\
&= \sum_{x,y,w} \mathbb{P}(X)\mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z)\mathbb{P}(W = w \mid PA) \\
&= \mathbb{P}(PA = z) \sum_{x} \mathbb{P}(X = x) \sum_{y} \mathbb{P}(Y = y \mid PA_y) \sum_{w} \mathbb{P}(W = w \mid PA)
\end{aligned}$$

$$= \mathbb{P}(PA = z)$$

Also, we have the following conditional distribution;

$$\begin{aligned}
& \mathbb{P}_m(Y = y \mid X = x, PA = z) \\
= \; & \frac{\mathbb{P}_m(Y = y, X = x, PA = z)}{\mathbb{P}_m(X = x, PA = z)} \\
= \; & \frac{\sum_w \mathbb{P}_m(Y = y, X = x, PA = z, W = w)}{\sum_{y,w} \mathbb{P}_m(Y = y, X = x, PA = z, W = w)} \\
= \; & \frac{\sum_w \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(X = x)\mathbb{P}(PA = z, W = w)}{\sum_{y,w} \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(X = x)\mathbb{P}(PA = z, W = w)} \\
= \; & \frac{\mathbb{P}(X = x)\sum_w \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z, W = w)}{\mathbb{P}(X = x)\sum_{y,w} \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z, W = w)} \\
= \; & \frac{\mathbb{P}(X = x \mid PA = z)\sum_w \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z, W = w)}{\mathbb{P}(X = x \mid PA = z)\sum_{y,w} \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(PA = z, W = w)} \\
= \; & \frac{\sum_w \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(X = x \mid PA = z)\mathbb{P}(PA = z, W = w)}{\sum_{y,w} \mathbb{P}(Y = y \mid PA_y)\mathbb{P}(X = x \mid PA = z)\mathbb{P}(PA = z, W = w)} \\
= \; & \frac{\sum_w \mathbb{P}(Y = y, X = x, PA = z, W = w)}{\sum_{y,w} \mathbb{P}(Y = y, X = x, PA = z, W = w)} \quad \text{using the factorization for } \mathbb{P} \\
= \; & \mathbb{P}(Y = y \mid X = x, PA = z)
\end{aligned}$$

Also note that, since PA consists of all the parent nodes of the nodes in X, it is evident that X and PA are d-separated in the modified DAG after applying the do-operator. This is because, the only way x and PA could not be d-separated if there is a path with chains or forks with no collider in the path, in the modified model. Also, X cannot have an incoming arrow in the modified model, since we are intervening on X which requires all incoming arrows to be removed. Therefore, all arrows (or edges) must come out of X, which tells that the series of dependence from X to PA must be carried throughout a chain. Hence, in the original model, there is a direct edge from PA to X, and there is a directed path from X to PA, hence, we have a directed cycle in the original DAG, which contradicts the definition of DAG (as it must be acyclic).

Therefore, we have $\mathbb{P}_m(PA = z \mid X = x) = \mathbb{P}_m(PA = z \mid X = x)$. Now, the rest of the proof follows from similar to the equation 1.

□

7.2 Backdoor Criterion

We know by means of **Adjustment Formula** that to find the effect on a DAG after intervening on X, can be estimated by adjusting the conditional distribution for all parent nodes of X. However, there are situations where the underlying causal DAG is obtained through means of scienfic laws, rahter than being estimated from the data alone. In such a case, the parent nodes of the intervening variable X might be immeasurable, or unobserved in the dataset. For instance, we refer back to Example 11, and change some of the variables. Let, X, Y, Z, W be four variables in the setup, where X is the indicator whether drug is given or not, Y denotes the recovery of the patient, Z denotes socio-economic status and W denotes the weight of the patient. Note that, socio-economic

status might influence whether the patient is able to visit a doctor and get the drug, and also it might influence the weight of the patient because of different food consumption patterns. Therefore, we would expect a DAG like Figure 10.

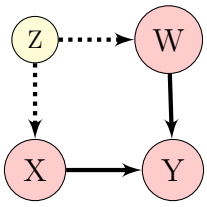


Figure 10: DAG for new drug testing setup with unobserved socio-economic status

Now, similar to Example 11, we wish to estimate the causal effect $\mathbb{P}(Y = y \mid do(X = x))$. However, in practical situations, hospitals do not record the socio-economic status of a patient, but the variable W, the patient's weight might be available from hospitals' registers. However, **Adjustment formula** suggests that we need to adjust for the parent of X, i.e. the variable Z, but we cannot apply adjustment formula as Z is not available. In such a situation, we need to find some other set of variables, for which we can perform the adjustment and it leads to the same estimate.

Before exploring how we can provide such an adjustment, let us focus on classifying how dependence can pass from X to Y.

1. There is a directed path from X to Y, i.e. a series of chains. In such a case, any change in X, in the modified DAG would result in changes in Y through this path.

2. There is a confounder variable from which a fork is created, and one end of the fork passes dependence to X, and another end passes dependence to Y. But, in the modified DAG, such a path would vanish, as we remove all incoming edges to X when intervening on Y.

Therefore, we can perform the conditioning on Z, and use it as an adjustment set, if;

1. Z does not block any paths of Type 1.

2. Z blocks all paths of Type 2.

3. Conditioning on Z does not create a path of Type 2. Note that, conditioning on a collider or a descendant of collider opens a path, hence this criteria effectively guards against creation of such spurious paths.

Definition 2 (Backdoor Criterion)**.** *Given an ordered pair of variables (X, Y) in a DAG G, a set of variables Z satisfies the* ***Backdoor Criterion*** *relative to (X, Y) if,*

1. *No node in Z is a descendant of X, and*

2. *Z blocks every path between X and Y that contains an arrow into X.*

To understand this definition, let us consider a set of variables Z, satisfying backdoor criterion. Note that, Z blocks every path that contains an arrow into X. This ensures that, every path between X and Y which essentially disappears after intervention, which has spurious relationship, is blocked. Also, since no node in Z is descendant of X, this

guards against the possibility that directed paths from X to Y are being blocked. Also, conditioning on descendants of collider is also being prevented by this criteria, since any descendant of collider between X and Y would also become a descendant of X.

In the above example of drug testing, note that W satisfies the backdoor criterion relative to (X, Y), and hence we could have an adjustment formula;

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_{w} \mathbb{P}(Y = y \mid X = x, W = w)\mathbb{P}(W = w)$$

From the discussion above, we could infer about the following theorem.

Theorem 2 (The Backdoor Adjustment Formula)**.** *If Z satisfy Backdoor criterion relative to (X, Y), we can estimate the intervening causal effect using adjustment with Z, i.e.*

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_{z} \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z)$$

Proof. Let PA denotes the set of parent nodes corresponding to the variables X. We then know because of Theorem 1,

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_{t} \mathbb{P}(Y = y \mid X = x, PA = t)\mathbb{P}(PA = t)$$

Now, we shall prove the following two lemma.

(i) $Y \perp\!\!\!\perp PA \mid (Z, X)$

(ii) $X \perp\!\!\!\perp Z \mid PA$

To show (i), consider a path between PA and Y,

1. The path passes through X, hence, that path would be blocked once we condition on X.

2. The path does not pass through X. In that case, we extend the path from PA to some node X, which creates a path from X to Y with an arrow into X from PA. Such a path would be blocked by Z as it satisfies backdoor criterion.

Therefore, Y and PA are d-separated with respect to (Z, X).

To show (ii), note that as Z does not contain any descendant of X, hence any path between Z and X, must have an arrow into X, which is from a parent node of X. Hence, PA blocks every path between X and Z.

Now, using the adjustment formula for parent nodes, we may write;

$$\begin{aligned}
& \mathbb{P}(Y = y \mid do(X = x)) \\
= & \sum_{t} \mathbb{P}(Y = y \mid X = x, PA = t)\mathbb{P}(PA = t) \\
= & \sum_{t}\sum_{z} \mathbb{P}(Y = y \mid X = x, PA = t, Z = z)\mathbb{P}(Z = z \mid X = x, PA = t)\mathbb{P}(PA = t) \\
= & \sum_{t}\sum_{z} \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z \mid X = x, PA = t)\mathbb{P}(PA = t) \quad \text{, by (i)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_t \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z \mid PA = t)\mathbb{P}(PA = t) \quad \text{, by (ii)} \\
&= \sum_z \mathbb{P}(Y = y \mid X = x, Z = z) \sum_t \mathbb{P}(Z = z \mid PA = t)\mathbb{P}(PA = t) \\
&= \sum_z \mathbb{P}(Y = y \mid X = x, Z = z) \sum_t \mathbb{P}(Z = z, PA = t) \\
&= \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z)
\end{aligned}
$$

This completes the proof. □

7.3 Frontdoor Criterion

There is still a lot of debate about the topic on whether Smoking causes lung cancer. Around 1970, the tobacco industry popularizes the theory that there is a unobserved variable which affects both the lung cancer and the smoking. The theory focuses on the idea that there is a particular genotype, that makes people attracted to tobacco and smoking, while the same genotype makes people prone to lung cancer. Therefore, we would get a causal DAG shown in Figure 11 where X denotes whether the person smokes or not, Y denotes whether the person has lung cancer or not, U denotes the particular unobservable genotype.

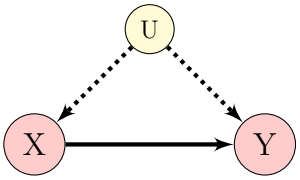


Figure 11: Tobacco industries' built up causal DAG

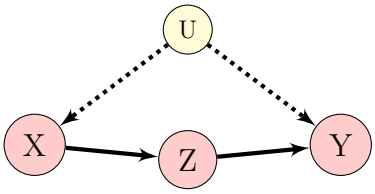


Figure 12: Causal DAG to counter tobacco industries' theory

Note that, due to such a causal DAG, the only possible backdoor path from X to Y is via U, hence U, the unobserved variable is the only variable satisfying backdoor criterion. Hence, if we wish to estimate the causal effect of X onto Y, by intervening on X, then it requires adjustment through the variable U, which is not possible due to its absence in the data. Hence, the causal effect is not estimable. And based on this theory, tobacco industry could survive for a long period.

But some recent study indicates that it is possible to measure a related variable, the tar deposits in lungs (through X-ray, Medical Imaging etc.) which is indicated as Z in Figure 12. It is clear that, Smoking may have an affect on the tar deposits in lung, and that might cause Lung cancer. However, it is obvious that due to presence of certain genes, it is not possible a person has tar deposits in his/her lungs by birth. Therefore, a causal DAG like shown in Figure 12 brings out an even clearer picture of the causation diagram.

However, even in the DAG given in Figure 12, the only backdoor variable is U, hence we cannot use the backdoor adjustment formula right away. But, repeated use of backdoor formula will help us estimating the causal effect, as we shall see.

We focus on the DAG in Figure 12. Firstly, note that, there is no backdoor path from X to Z, hence the empty set ϕ, actually satisfies the backdoor criterion. Hence, using backdoor adjustment formula, we get;

$$\mathbb{P}(Z = z \mid do(X = x)) = \mathbb{P}(Z = z \mid X = x)$$

Now that, we have a grip on the effect from X to Z, let us try to analyze the effect from Z to Y. Note that, if we intervene on Z, then X being a parent of Z satisfies backdoor criteria. Hence,

$$\mathbb{P}(Y = y \mid do(Z = z)) = \sum_x \mathbb{P}(Y = y \mid Z = z, X = x)\mathbb{P}(X = x)$$

Therefore,

$$\begin{aligned}
& \mathbb{P}(Y = y \mid do(X = x)) \\
= \quad & \sum_z \mathbb{P}(Y = y \mid do(X = x), Z = z)\mathbb{P}(Z = z \mid do(X = x)) \\
= \quad & \sum_z \mathbb{P}(Y = y \mid do(Z = z))\mathbb{P}(Z = z \mid do(X = x)) \\
& \qquad \text{since Z is descendant of X, intervening on X and conditioning on Z} \\
& \qquad \text{is same as intervening on Z} \\
= \quad & \sum_z \sum_{x'} \mathbb{P}(Y = y \mid Z = z, X = x')\mathbb{P}(X = x')\mathbb{P}(Z = z \mid X = x)
\end{aligned}$$

This formula obtained from repeated usage of backdoor is called **Frontdoor formula**. This is possibly because of the reason that it is based on adjusting for a variable lying on a direct path from X to Y.

Definition 3 (Frontdoor Criterion)**.** *A set of variables Z is said to satisfy the **Frontdoor Criterion** relative to an ordered pair of variables (X, Y) if;*

1. *Z intercepts all directed paths from X to Y.*
2. *There is no unblocked path from X to Z.*
3. *All backdoor paths from Z to Y are blocked by X.*

Note that, the set of all children of X satisfy frontdoor criteria. The 1st condition basically indicates that applying intervention on X and applying conditioning on Z, would effectively turn into intervening on Z. The 2nd condition essentially implies that the first backdoor criterion applies with empty set ϕ. The third condition states that the second backdoor criterion applies with the intervention variable X.

Theorem 3. *If Z satisfy frontdoor criterion relative to (X, Y), and if $\mathbb{P}(X = x, Z = z) > 0$, i.e. we can condition on this event, then;*

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_z \mathbb{P}(Z = z \mid X = x) \sum_{x'} \mathbb{P}(Y = y \mid Z = z, X = x')\mathbb{P}(X = x')$$

Proof. The proof basically follows from the interpretation of the definition of frontdoor criterion and the technique similar to derivation of frontdoor adjustment formula for causal DAG given in Figure 12. □

7.4 Inverse Probability Weighting

Consider the adjustment formula with adjustment variable being Z,

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z)$$

Observe that, in the right hand side of the equation, the summation is over all possible values of the variable Z. However, if there is a large number of adjustment variables, since we have no control over the size of adjustment set (the set consisting of all adjustment variables), the number of summands increases exponentially. For instance, if we have 10 binary variables in the adjustment set, then there will be $2^{10} = 1024$ terms to sum in the right hand side of the adjustment equation. However, most of those terms would be estimated to be 0, which may not be obvious at first. However, we can rewrite the adjustment equation as;

$$\begin{aligned}
\mathbb{P}(Y = y \mid do(X = x)) &= \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z) \\
&= \sum_z \frac{\mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z)\mathbb{P}(X = x \mid Z = z)}{\mathbb{P}(X = x \mid Z = z)} \\
&= \sum_z \frac{\mathbb{P}(X = x, Y = y, Z = z)}{\mathbb{P}(X = x \mid Z = z)}
\end{aligned}$$

Note that, the above sum only contributes something nonzero, if and only if we have a sample datapoint with $X = x, Y = y$ and $Z = z$. Hence, the above sum can actually be replaced with the sum over all samples in the dataset. Hence,

$$\hat{\mathbb{P}}(Y = y \mid do(X = x)) = \frac{1}{\text{Total number of samples}} \sum_{\text{samples}} \frac{1}{\mathbb{P}(X \mid Z)}$$

which is a weighted sum of the samples, where each sample is weighted with $\frac{1}{\mathbb{P}(X|Z)}$.

7.5 Mediation

Often it is the case that, the dependence from one variable to another variable is passed through a direct edge, and also through a series of chains. Then, we would have a direct effect passed through the direct edge, and also some mediating effect through a series of chains, i.e. through some other variables which passes the dependence.

For instance, suppose we wish to know whether and to what degree a company discriminates by gender (X) and its direct hiring practices (Y). However, qualifications Q may be a mediating variable to transfer dependence from gender to hiring practices. Therefore, a representing DAG to describe the above causality structure is like Figure 13.

Now, in the causal DAG given in Figure 13, if we want to estimate the direct effect from G and H, then we can perform adjustment on the variable Q. However, if there is a confounder between the mediating variable Q and the response H, then the situation changes entirely. Here, Socio-economic status I might be such a confounder between Q and H, i.e. people with high income or from higher socio-economic background can afford to go for higher education and obtain higher qualifications, and also being from higher socio-economic background might create some connections which can help one to

get hired. The corresponding causal DAG is shown in Figure 14. Note that, under this new causal DAG, if we try to estimate the direct effect from G on H, then we have two situations:

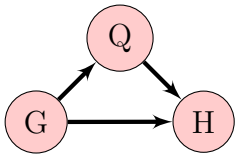


Figure 13: Causality DAG for relationship between Gender, Qualifications and Hiring

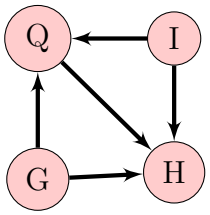


Figure 14: Causality DAG for relationship between Gender, Qualifications, Hiring status and Income states

1. If we condition on Q, then we are essentially conditioning on a collider, as then the dependence would go through the path Gender $\rightarrow$ Qualifications $\rightarrow$ Income $\rightarrow$ Hiring Status.

2. If we do not condition on Q, then the dependence will pass through the chain; Gender $\rightarrow$ Qualifications $\rightarrow$ Hiring Status.

So, as we can see that there is no way through simple conditioning which would allow one to estimate direct effect from G to H. Luckily, there is a conceptual way of holding the mediating variable steady without conditioning on it: We can intervene on it. So, instead of conditioning, if we fix the qualifications, the arrow between gender and qualifications (and the one between income and qualifications) disappears, and no spurious dependence can pass through it.

Definition 4. *If X, Y and Z be any three variable in a causal DAG with Z being a mediator between X and Y, i.e. there is a chain $X \rightarrow Z \rightarrow Y$ and there is an edge from X to Y, then the* ***Controlled Direct Effect (CDE)*** *on Y of changing the value to X from x' to x is given as;*

$$CDE(z) = \mathbb{P}(Y = y \mid do(X = x), do(Z = z)) - \mathbb{P}(Y = y \mid do(X = x'), do(Z = z))$$

The obvious advantage of this definition over the one based on conditioning is its generality; it captures the intent of "keeping Z constant" even in cases where the $Z \rightarrow Y$ relationship is confounded. Note that the direct effect may differ for different values of Z; for instance, it may be that hiring practices discriminate against women in jobs with high qualification requirements, but they discriminate against men in jobs with low qualifications. Therefore, to get the full picture of the direct effect, we'll have to perform the calculation for every relevant value z of Z. However, in case of linear functional relationship between two variables, such direct effect is basically the slope of the linear relation multiplied by the difference $(x - x')$, which remains constant for each value z to Z.

Now, to estimate this controlled direct effect, we need to estimate terms like $\mathbb{P}(Y = y \mid do(X = x), do(Z = z))$. The estimation procedure of an expression containing more

than 1 variable being intervened on, is similar to that of usual procedure of estimation of quantities with a single $do(\cdot)$ operator, just proceeding sequentially reducing do operators one at a time. First note that, according to Figure 14, there is no backdoor from G to H, hence the do operator applied on G can simply be replaced with simple conditional distribution. Hence,

$$\mathbb{P}(H = h \mid do(G = g), do(Q = q)) = \mathbb{P}(H = h \mid G = g, do(Q = q))$$

Now, we note that there is only one backdoor path from Q to H, and this is through the fork with root at I. Hence, adjusting for I would help us to remove the remaining do operator, hence we finally obtain;

$$\mathbb{P}(H = h \mid do(G = g), do(Q = q)) = \sum_i \mathbb{P}(H = h \mid G = g, Q = q, I = i)\mathbb{P}(I = i)$$

This last formula is do-free, hence can be estimated from the data, if we have data on the variable I. In general, the CDE of X on Y, mediated by Z, is identifiable if the following two properties hold:

1. There exists a set S_1 of variables that blocks all backdoor paths from Z to Y.
2. There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z.

Note that, the first condition ensures that we can remove the do operator for Z, and the second condition ensures that we can remove the do operator for X, which finally gives an estimate of CDE based on adjusting for S_1 and S_2.

It is even trickier to determine the indirect effect than the direct effect, because there is simply no way to condition away the direct effect of X on Y. It's easy enough to find the total effect and the direct effect, so some may argue that the indirect effect should just be the difference between those two. This may be true in linear systems, but in nonlinear systems, differences don't mean much; the change in Y might, for instance, depend on some interaction between X and Z, if, as in the previous example, women are discriminated against in high-qualification jobs and men in low-qualification jobs, subtracting the direct effect from the total effect would tell us very little about the effect of gender on hiring as mediated by qualifications. Clearly, we need a definition of indirect effect that does not depend on the total or direct effects. Counterfactuals may be used to truely estimate this type of indirect effects.

8 Applications on Real Life Data

8.1 Effectiveness of Job Training Programme

In 1986, Robert J. LaLonde, in his paper titled **Evaluating the Econometric Evaluations of Training Programs with Experimental Data**, collected an experimental data with the help of contemporary Government agency called **Panel Study of Income Dynamics (PSID)**. In his study, he concentrated on a group of 445 individuals, with 10 socio-economic variables. The 10 variables available are as follows:

1. `Age` - Agegroup, starting from 16 to 5 years of interval.
2. `Education` - Uneducated, Primary, Secondary, Higher secondary, Graduate, Post graduate based on number of years of schooling.
3. `Black` - 1 if black, 0 otherwise.
4. `Hispanic` - 1 if Hispanic, 0 otherwise.
5. `Married` - 1 if married, 0 otherwise.
6. `Nodegree` - 1 if no high school degree, 0 otherwise.
7. `re74` - Income group (Unemployed, Low, Medium, High, Very high) in 1974.
8. `re75` - Income group (Unemployed, Low, Medium, High, Very high) in 1975.
9. `re78` - Income group (Unemployed, Low, Medium, High, Very high) in 1978.
10. `Treatment` - 1 if the individual receives the Job training program in 1974, 0 otherwise.

On this dataset, Lalonde performs his method of estimating effects of Job training program, and determines that the Job training program actually helps people to increase income in subsequent years.

However, we can use the d-separation to search for a Causal DAG model which can accurately describe the data generating mechanism. Then, we get a DAG model which looks like Figure 15. Note that, the DAG model says that there is no other variables other than Wage in 1975 and the treatment (i.e. the Job training program) has effect on Wage in 1978. Note that, there is association between Black and Hispanic, between Married and AgeGroup, between Education and Nodegree, which actually makes sense. Also, time dependence between Wages in 1974, 1975 and 1978 are clearly depicted. However, such a nice structure of causal DAG actually tells that there is only a direct effect of Job training to Real Wage in 1978, which can be estimated without any sort of adjustment.

Much later after Lalonde's experiment, PSID publicly released all of the observational data with the same socio-economic variables on 2675 individuals under its observation. Lalonde specifically used a subset of only 445 among these. Applying the causal model searching technique as before, we get a causal DAG as in Figure 16. Such a causal DAG was estimated without imposing any restriction of association between two variables. Note that, Married and Black and Real Wage in 1974, these three variables has an arrow into the treatment variable, thereby indicating the fact that the treatment was not possibly randomly distributed. It might be possible that the Job training program was more taken by Unemployed people and it was bound to show some improvement after three years. Also note that, being Black affects the Wage in 1978. However, there is one contradictory arrow in the Figure, going from Education to Black, which should be reverse in direction by common intuition.

The main thing to notice that the treatment variable does not have an arrow to Wage in 1978, hence from the data, we do not find any direct causal relationship between Job training program and Wage in 1978. However, we can perform the Causal model searching in a restricted space where we consider an edge between Treatment and Re78 variables. The estimated causal DAG under this restricted search space is shown in Figure 17. But

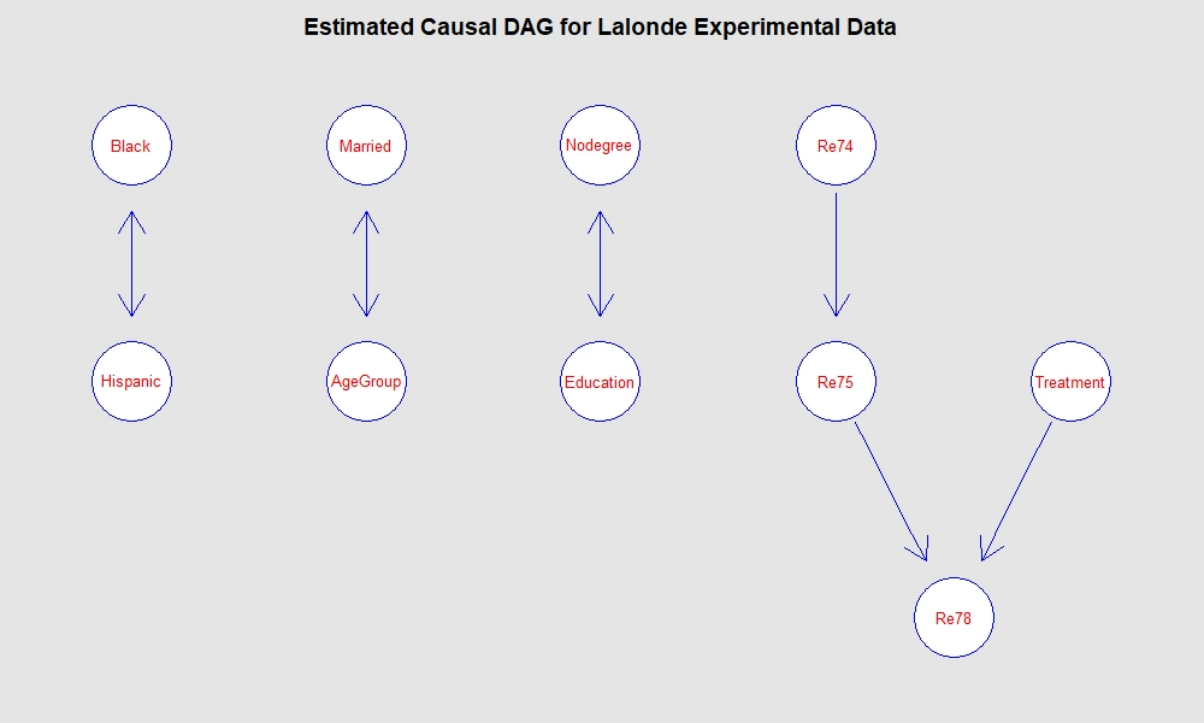


Figure 15: Estimated causal DAG based on Lalonde's data, 1986

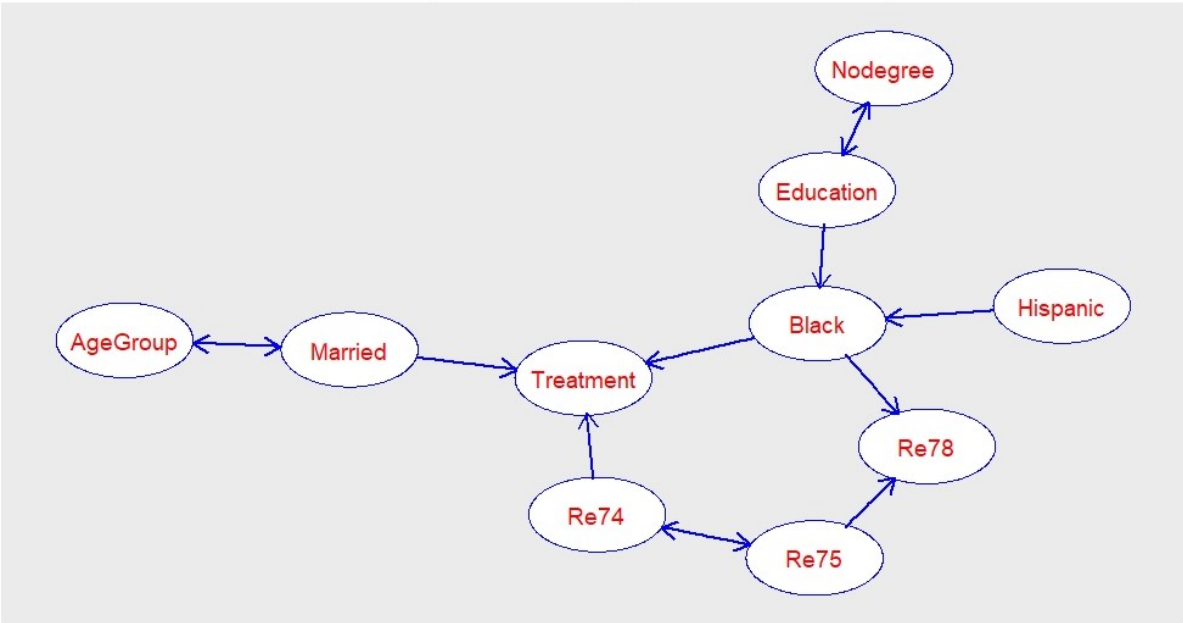


Figure 16: Estimated causal DAG based on PSID data

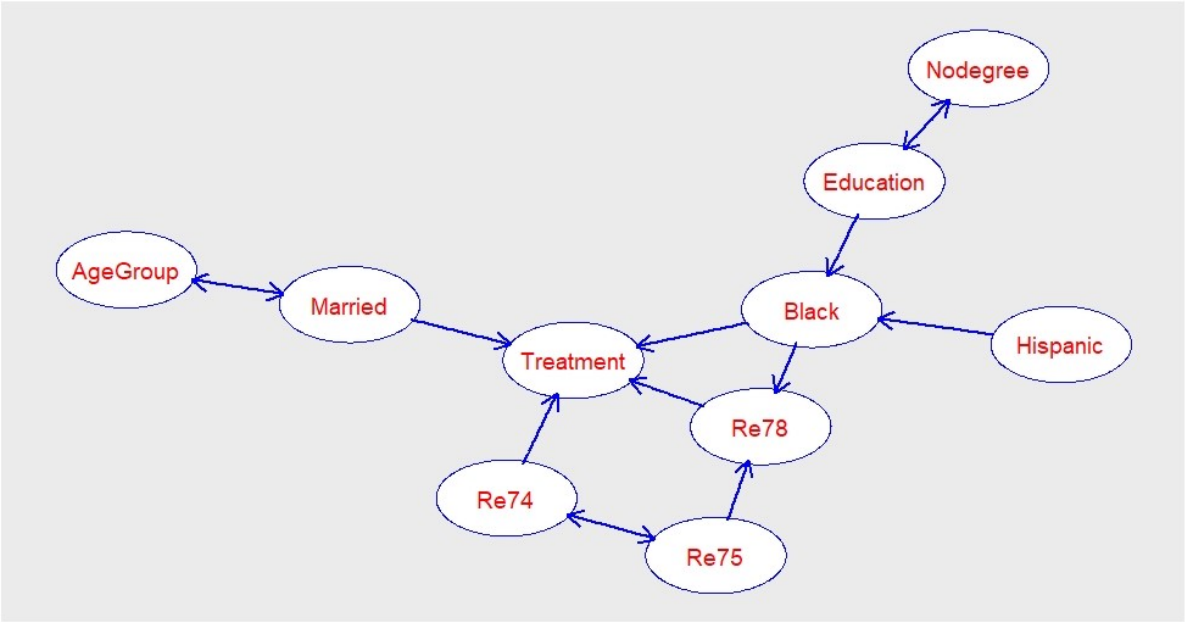


Figure 17: Estimated causal DAG based on PSID data with restriction on association between Treatment and Re78

notice that, the arrow is estimated to be from Re78 to Treatment. But the treatment i.e. the Job training program happened in $1974 - 75$, while Re78 is a variable generated and measured at 1978. Hence, our imposed restriction creates false implications of causality. However, in any case, contrast to Lalonde's own findings, we find that all of the PSID data actually presents no relationship between Job training program and Real wages in 1978, hence signifying that increase in real wages in 1978 is not particularly caused by Job training program. It also shows that the spurious association between Treatment and Re78 is found only due to the fork with Re74 at its root.

8.2 Ability and Intelligence Tests

The **Ability and Intelligence Tests** data available in `datasets` package in `R`, contains data on 6 tests given to 112 individuals. This dataset is popularly used for Factor Analysis or Latent Variable Analysis. However, we wish to determine which variable actually affects the other and what is a causality diagram for these skills related to these variables. This will help one to understand better how enhancing one skill through training would also increase other types of skills.

The tests given to the individuals are as follows:

1. `General`, A non-verbal measure of general intelligence using Cattell's culture-fair test. It is popularly used as a basic measure of IQ tests. It requires a person to identify a pattern of images and complete that pattern.

2. `Picture`, A test where an individual is given pieces of an image, and he/she is required to find the missing piece to complete the picture.

3. `Blocks`, A test where the test taker need to rotate painted blocks or pieces in position to complete a design.

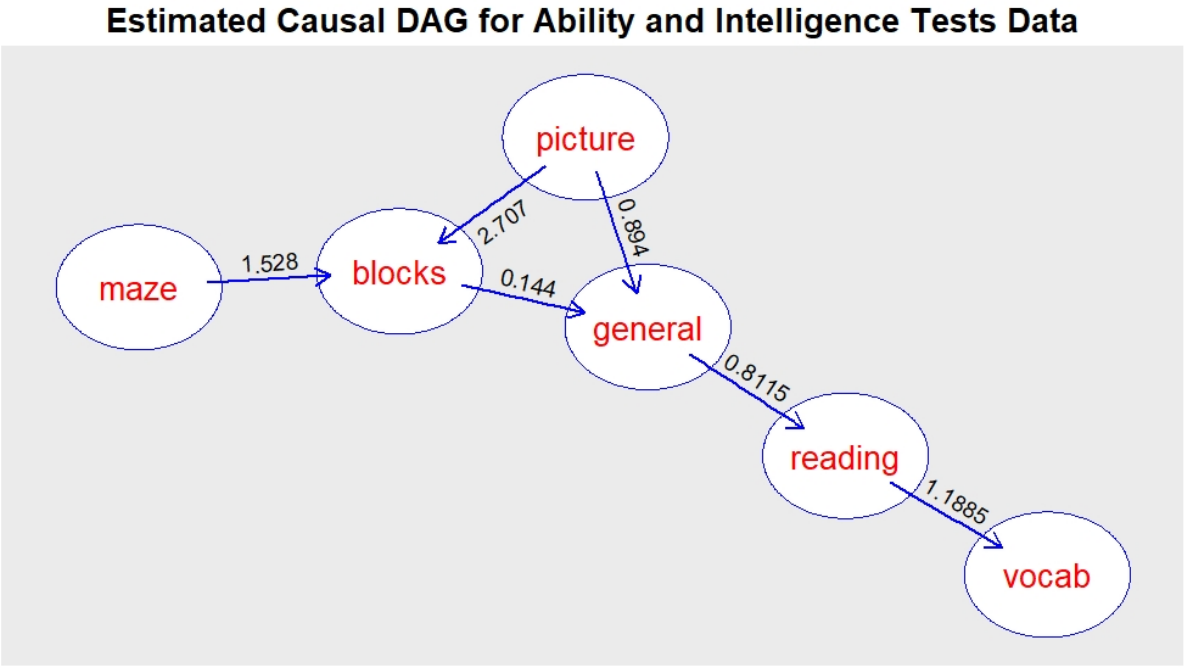


Figure 18: Estimated causal DAG for Ability and Intelligence Tests dataset

4. `Maze`, A test where given a starting point and an endpoint, you have to go from start to end through a maze of varying complexity.

5. `Reading`, A reading comprehension test.

6. `Vocab`, A vocabulary test.

Since all of scores are continuous, they are centered at 0, and a linear relationship is considered through each of the causality arrow. As both variables are the endpoint in arrows are centered and have mean 0, hence only one number, determines their relationship. For instance, if we have an edge from X to Y in the causal DAG, then we consider the relationship as $Y = \beta_{0,yx} + \beta_{1,yx}X$. However, due to centering only $\beta_{1,yx}$'s are estimated, and $\beta_{0,yx}$'s can simply be estimated from adjusting for the corresponding means, with $\hat{\beta}_{0,yx} = \bar{Y} - \hat{\beta}_{1,yx}\bar{X}$. Writing these estimates of slope on each corresponding edges, we obtain the estimated causal Graph as shown in Figure 18. Now, for instance, if we want to know the effect that playing Maze game would help us in building Vocabulary, we compute its direct effect by multiplying corresponding coefficients in each paths and add contribution of all paths; $1.528 \times (0.144 + 0.894/2.707) \times 0.8115 \times 1.1885 = 0.698$, which is something different from usual intuition.

8.3 Determination of Prices of Tea in Auctions through India

In India, e-auctions occur pan India since September 2016. This means, that any registered tea seller can submit bids for the tea packets. These auctions occur between the warehouse, and the potential sellers in the market. The garden to auction mechanism acts as follows: Manufactured tea is dispatched from various gardens/ estates to the auction centres for sale through the appointed auctioneers, on receipt of which, the warehouse keeper sends an arrival an 'weighment report' showing the date of arrival and other details

pertaining to the tea including any damage or short receipt from the carriers. The tea is catalogued on the basis of their arrival dates within the framework of the respective Tea Trade Associations, the quantities are determined according to the rate of arrivals at a particular auction centres. Registered buyers, representing both the domestic trade and exporters receive samples of each lot of teas catalogued, which is normally distributed a week ahead of each sale enabling the buyers to taste, inform their principals and receive their orders well in time for sale. On the other hand, auctioneers themselves taste the tea, and give an 'educated guess' at what price the tea should be sold at, on the basis of which, the base price of the auction is set. Then the tea auctions are held online, and the highest bidder wins, who pays a price equal to the second highest bid.

Note that this procedure has the only formal subjective procedure of the auctioneers' valuation. The degrading quality of tea, which had been a long standing concern for the corresponding sellers and consumers, could be associated to this subjectivity of the auctioneers. We had tried a simple linear model of the price on the valuation of auctioneers, and other relevant variables (source of tea packet, tea grade, category of tea, supply of tea in that week and the month of the year) from the J. Thomas' dataset on the year of 2018, which led to a R^2 of 0.92, and significance of all variables.

However, we could not yet allege that the auctioneers' valuation had a causal impact on price level, although it seemed to be an indispensable predictor. To evaluate the causal impact of valuation on price, we lay down a part of the causal graph structure in Figure 19. Note that there may be some causation structure between the variables (for example, not all grades occur in all months, hence there should be an arrow from month to grade). Nevertheless, the variables of interest are the valuation and the final price, and the parents to them known, assuming we subscribe to a linear understanding of time and causality. Thus all the other variables in the model are parents to both valuation and price. Furthermore, valuation could be a parent of price. But, if valuation is indeed an 'educated guess' of price as was originally intended, then, since the guess is based on only these variables; conditioned on these variables, price and valuation should be independent trials from (possibly same) distribution. Hence, they should be independent, had valuation not been a cause of price. Hence we aim to test this causality.

Thus, we wish to test whether

$$\text{Valuation} \perp\!\!\!\perp \text{Price} \mid (\underbrace{\text{Source, Grade, Volume, Garden, Month}}_{\text{rest}})?$$

where the residuals are However, we are stumped by the fact that conditioning on so many variables render too less data to provide reliable estimates for any inference. Hence, we take a different strategy, which is generally often taken when the conditioning variables are continuous. We model the log of valuations by a linear function of the other variables, and similarly for log of price. That is

$$\begin{aligned} \Omega_1 : \log \text{Valuation} &= \beta_1^\intercal \text{rest} + \varepsilon_1, \qquad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \\ \Omega_1 : \log \text{Price} &= \beta_2^\intercal \text{rest} + \varepsilon_2 \qquad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2), \varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \end{aligned}$$

If these models provide good fit, then we can find out the correlation between the residuals helps to identify the presence of the causal link between valuation and price. This is because, for the residuals in both the models are independent of rest, and for an two variables A, B and C, we have,

$$A \perp\!\!\!\perp B|C \Leftrightarrow [A - f(C)] \perp\!\!\!\perp [B - g(C)]|C \Leftrightarrow A - f(C) \perp\!\!\!\perp B - g(C)$$

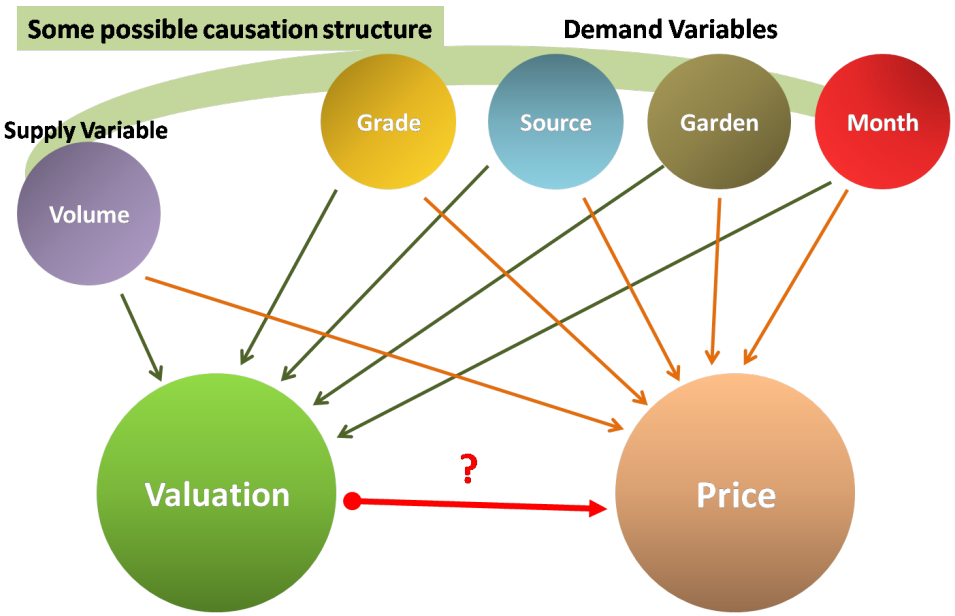


Figure 19: Causal Diagram for modeling Tea auction process

if $A - f(C)$ and $B - g(C)$ are independent of C itself. This is akin to testing for the partial correlation between valuation and price to be 0.

The results that we obtain from the data are as follows:

- Ω_1 and Ω_2 provides reasonably good fit, yielding the value of the multiple R^2 to be 0.75 and 0.71 respectively. The regression diagnostics checks were satisfied.
- The correlation between the residuals of Ω_1 and Ω_2 came out to be a whooping 0.859; which cannot be attributed to sheer chance with such a large sample size.

This has widespread implications. First of all, this substantiates that valuation of the auctioneers are not as harmless as providing an educated guess for the price, but rather have a causal impact on the final price. This occurs, as the base price is visible to all potential buyers. Thus, to automate the process, whatever procedure we propose cannot be held against the standard with how the valuation predicts price, as the predictability is nested as a causal impact. Hence, to automate the process, a possible change in the auction mechanism is required, and the premises of checking the success of any such alternate mechanism with this data (where valuation has had a causal impact) would be flawed.

9 Conclusion

This report provides a basic review about Causal Inference using Direct Acylic Graphs, and laids out simple mathematical foundations of do-calculus. It discusses how causal inference as a separate topic can be very powerful to provide answers to questions which statistics, as an individual subject cannot answer. It discusses about how d-seperation and conditional independences in a DAG model, helps us to search for a underlying causal structure which adequately explains the data generating mechanism. We also discuss about do-calculus, adjustment formula, backdoor and frontdoor criterion, which actually enables one to make inferences on a experimental setup, by a mathematical

tool of intervention in a non-experimental data. Finally, we perform some analysis of causal inference on different datasets, to estimate the underlying causal model, as well as estimating causal effects between two variables. In particular, we have noted some significant implications for Tea auction mechanism in India, using tools we have built in this report.

10 Acknowledgements

We would like to extend our thanks to Prof. Soumendu Sundar Mukherjee, who has introduced this interesting topic to us and suggested some good resources to help us learn this topic.

References

[1] *An Introduction to Causal Inference*, Judea Pearl, 2010, The International Journal of Biostatistics, Volume 6, Issue 2, Article 7.

[2] *Video Lectures on Causal Inference* Jason A Roy, A Crash Course in Causality: Inferring Causal Effects from Observational Data; *Coursera*

[3] *Causal Inference in Statistics: A Primer*, Judea Pearl, Madelyn Glymour, Nicholas P. Jewell. Wiley, 2016.

[4] *Evaluating the Econometric Evaluations of Training Programs.*, LaLonde, Robert. 1986. American Economic Review 76:604-620.

[5] *Latent Variable Analysis and Factor Analysis*, Bartholomew, D. J. (1987). Griffin.

[6] *Feasibility of Transparent Price Discovery in Tea through Auction in India* D. Mukherjee, A. Dalal, S. Roy (2019). MCX Commodity Insights Yearbook.