



UNIVERSITÁ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE
ED AZIENDALI "M. FANNO"

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

CAUSAL DIAGRAMS FOR CAUSAL
INFERENCE: AN INTRODUCTION

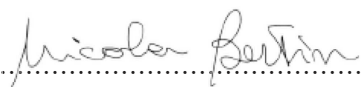
RELATORE:
CH.MO PROF.RE RETTORE ENRICO

LAUREANDO: BERTIN NICOLA
MATRICOLA N. 1220622

ANNO ACCADEMICO 2021-2022

Dichiaro di aver preso visione del “Regolamento antiplagio” approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione ‘Riferimenti bibliografici’.

I hereby declare that I have read and understood the “Anti-plagiarism rules and regulations” approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section ‘References’.

Firma (signature) 

English abstract

In the econometric-statistical field, to quantify a cause and effect relationship it is required a detailed preliminary work of knowledge of the context in which the studied phenomenon occurs, in order to provide useful arguments to attribute a causal interpretation to a correlation found empirically: "Correlation does not imply causation". The causal diagrams, more precisely the 'Directed Acyclic Graphs' (DAGs), are an effective tool to synthesize and communicate the system of causal relationships that occur in the context in which the causal inference analysis takes place and, therefore, to set up the research work. This paper aims to explain, with simple words and examples, why causal inference requires a preliminary knowledge of the context and, then, how to use the DAGs to set your own research in order to find the searched cause-effect relationship. The paper is thought for who approaches to this discipline with minimal statistical bases: simple and multiple regression; graphs.

Italian abstract

In ambito econometrico-statistico, per quantificare una relazione causa effetto è richiesto un dettagliato lavoro preliminare di conoscenza del contesto in cui si manifesta il fenomeno studiato, al fine di fornire argomenti utili ad attribuire una interpretazione causale ad una correlazione riscontrata empiricamente: "Correlazione non implica causalità". I diagrammi causali, più precisamente i 'Directed Acyclic Graphs' (DAGs), sono uno strumento efficace per sintetizzare e comunicare il sistema di relazioni causali che si presentano nel contesto in cui si svolge l'analisi di inferenza causale e, di conseguenza, per impostare il lavoro di ricerca. Questo elaborato mira a spiegare, con parole semplici ed esempi, perchè l'inferenza causale richiede una conoscenza preliminare del contesto e, poi, come utilizzare il DAGs per impostare la propria ricerca al fine di trovare la relazione causa-effetto ricercata. L'elaborato è pensato per chi si avvicina a questa disciplina con minime basi statistiche: regressione semplice e multipla; grafici.

Contents

1	Introduction	1
1.1	What is <i>Causal Inference</i> ?	1
1.2	What is a <i>Causal Research Question</i> ?	1
1.3	<i>Data Generating Process, Identification and Research Design</i>	2
1.4	Identification cannot be known	3
1.5	To sum up	3
1.6	Why is statistics not enough?	3
1.7	The Simpson's Paradox	4
1.8	And now <i>Causal Diagrams</i> come	5
2	Causal Diagrams: how are they made?	7
2.1	Causal Diagrams or <i>Directed Acyclic Graphs</i> ?	7
2.2	Basic elements: <i>Nodes</i> and <i>Arrows</i>	7
2.3	Causal Effects	8
2.4	Fundamental assumptions	9
2.5	Common structures	9
2.5.1	Confounders	9
2.5.2	Colliders	10
2.5.3	Moderators	11
3	How to draw Causal Diagrams?	15
4	How to use Causal Diagrams?	17
4.1	Paths: what are they?	17
4.2	How to find all the paths	18
4.3	Categorize the paths found	19
4.3.1	<i>Front door</i> and <i>Back door paths</i>	19
4.3.2	<i>Good</i> and <i>Bad paths</i>	20
4.4	Get identification: to close all the <i>bad</i> paths	21
4.4.1	<i>Open</i> and <i>Closed paths</i>	21

4.4.2	How to close a path? <i>Controlling for</i>	22
4.4.3	Pay attention to <i>Colliders</i>	24
4.4.4	Pay attention to <i>Moderators</i>	26
4.4.5	The <i>Backdoor Criterion</i>	27
4.4.6	To Sum Up	27
4.5	The <i>Placebo Test</i>	28
5	Not only closing backdoors	29
5.1	Exogenous source of variation	29
5.2	The <i>Front Door Method</i>	32
6	Conclusion	35
6.1	To Sum Up	35
6.2	In short, my experience	35
	Bibliography	37

Chapter 1

Introduction

1.1 What is *Causal Inference*?

Causal Inference is that branch of econometrics that deals with identifying and quantifying empirically - that is on the basis of observations collected from the real world [Hun21, p. 4] - relationships characterized by a *cause-effect relationship*.

1.2 What is a *Causal Research Question*?

Any causal inference problem begins with a *causal research question*, that is a question whose answer is a numerically quantifiable cause-effect relationship. For example: "How does the hourly wage vary as a result of an increase in the level of education?" [Car99] "Does the granting of a residence permit to an illegal immigrant on average reduce the likelihood of this person committing a crime?" [Pin17] "Does the introduction of a minimum wage have an impact on the unemployment rate?" [CK00] "Does the presence of organised crime in a certain Italian region have an impact on per capita GDP?" [Pin15]. We can identify two common elements that underlie each of these questions:

1. A variable whose variation is source of the causal effect and which we will call ***Determinant*** (then shortened **D**). The determinant is more commonly called ***Treatment***, since it often consists in subjecting or not a certain subject to a certain situation, intervention.
2. A variable that varies because of the causal effect subjected and that we will call ***Outcome*** (then shortened **Y**)

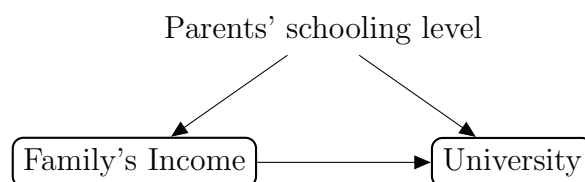
Therefore, defined these two elements, we have the possibility to give a general definition

of *causal effect* [Hun21, p. 89]:

Variable D has a causal effect on variable Y when a change in the value assumed by D unilaterally changes the value assumed by Y, ceteris paribus

1.3 *Data Generating Process, Identification and Research Design*

Our objective is to *identify* and quantify *just* the causal effect of D on Y. The *identification process* can be simple in *experimental environments*, where a scientist can control all the variables of the system in which he/she operates, but this is not the case in reality. In fact, it often happens that the value assumed by a certain variable is the result of several variables that affect the value of this one. For example, consider the causal effect of *family's income* on the *individual's probability of enrolling in academic studies* for a graduate student. The probability of enrolling in academic studies probably depends on other factors, such as the *individual's ability* or *individual's motivation*. Here we want an example as clear as possible, so we have inserted just another variable: the *parents' schooling level*. If we only want to evaluate the causal effect of *family's income* on *individual's probability of enrolling in academic studies*, how can we do it? In other words, how can we *identify* just the causal effect of *family's income*, getting rid of the causal effect generated by *parents' schooling level*¹ This example allows us to introduce two fundamental concepts: the *Data Generating Process* and the *identification*. The **Data Generating Process** (then shortened **DGP**) is *the set of underlying laws that determine how the data we observe are generated* [Hun21, p. 67]. **Identification**, instead, is that *situation in which it has been possible to isolate the searched causal effect, separating it from every other possible source of variation coming from the other factors that are present in our data generating process*. Going back to our example, the DGP will be the set of all the causal relations that exist in the system "individual's probability of enrolling in academic studies".



Since our goal is to recognize the effect of *family's income* on the *individual's probability of enrolling in academic studies*, we will have obtained identification when, following a certain procedure, we will have isolated the effect caused by *family's income* compared

¹If there were more variables the problem is the same: "How to get rid of the causal effect of all the variables except *family's income*?"

to the effect caused by *parent's schooling level*. The protocol we intend to follow to obtain identification, which will consist in the sequential application of different statistical-econometric tools, is called **Research Design**.

1.4 Identification cannot be known

What makes causal inference complex is the fact that we can never be certain we have achieved identification. An example is the best way to understand this concept. Let's suppose the true, but unknown, causal effect of *family's income* is such that every 50.000\$ the individual's probability of enrolling in academic studies increases by 5%. But we got 3% with our research design. So we didn't get identification: we're underestimating the real value. The problem is that there is no way to verify it, to know if we caught the real value. For this reason research design is fundamental, because according to this our result will acquire credibility. The more we see to research design the more our research answer will acquire *robustness*.

1.5 To sum up

We have a question whose answer consists in a causal relationship, i.e. we have a *causal research question*. We want to quantify this effect-causing relationship basing our research on real-world observations. In real world cause-effect relations tend not to be distinct and easily identifiable, since different variables interact with each other and jointly contribute to determine more cause-effect relationships, also affecting the causal effect that we are looking for. The set of all these relationships is called the *Data Generating Process*. Our goal is to identify only the cause effect relationship that we are interested in, eliminating any distortion from other variables in the DGP. In short, we want to *identify* our *causal effect*: to obtain *identification*. The set of econometric operations and tools that we intend to use to obtain identification is our *research design*.

1.6 Why is statistics not enough?

In causal inference the use of statistical tools, particularly the use of multiple regression, is not sufficient to identify the searched causal effect. Why? Because statistical tools can tell if there is *correlation (association)* between two variables, but not if there is a *causal relationship*. To give a definition to correlation and causation we could say that:

Correlation (also "Association"): two variables X and Y are correlated when they occur together following a certain relationship;

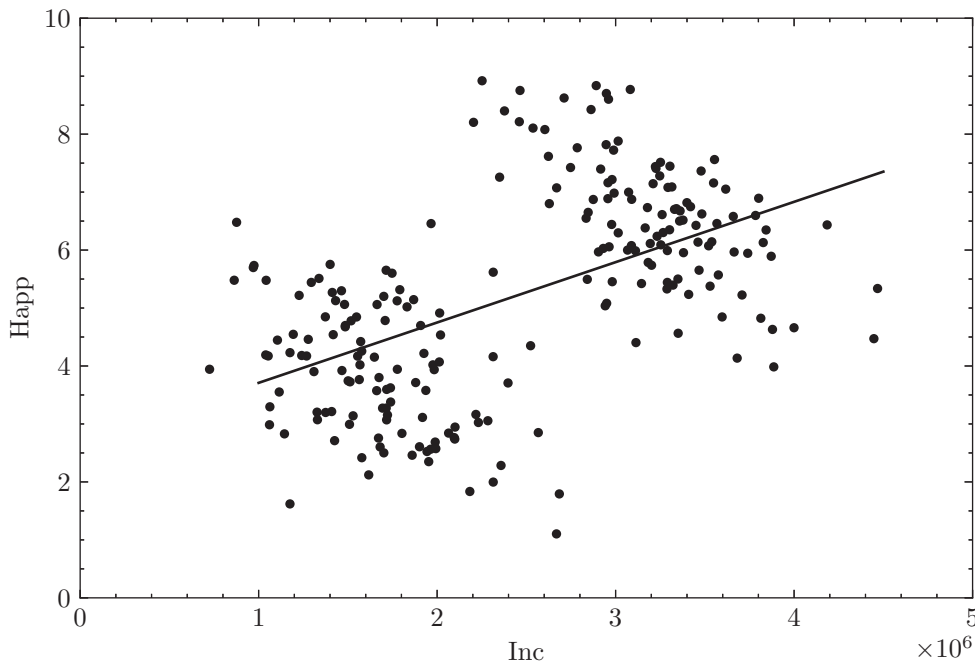
Causation: two variables X and Y are linked by a causal relationship when they are correlated and *we know that one causes the other*.

So we can say that “*correlation does not imply causation*” [Wikd] and this is the reason why we cannot rely only on statistical tools.

Suppose we have a variable Y *regressed*² on D. The regression coefficient will tell us how Y and D are related but, to say that D causes Y and that the regression coefficient represents the causal effect, we must first have assumed in the DGP that Y is caused by D.

1.7 The Simpson’s Paradox

Another reason why causal inference requires prior knowledge of DGP is given by the Simpson’s paradox, which says that ignoring even just one variable, the result could be strongly distorted. Let’s take an example. ”Does a higher income bring greater happiness to people?” [Cha21]. Collecting data on income and happiness and tracing a regression line we get the relationship, the *correlation*, is positive and, based on our assumptions, we say that income causes greater happiness.

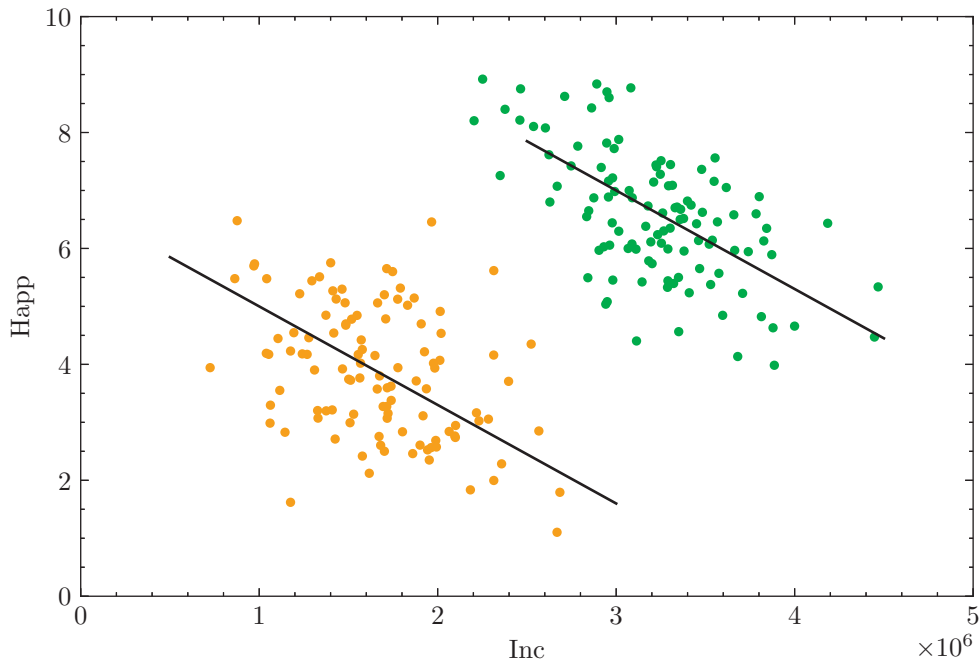


However, if we consider that the data come from two different countries, let’s suppose Canada (green points) and Mexico (orange points), where the average income is very different, we notice that in reality the relationship is negative³.

This is how, considering a variable that was previously hidden, the result changes. For

²That means there is a regression line like $Y = \beta_0 + \beta_1 D + \dots + \mathcal{E}$

³Thanks to Alessandro Miotto for providing me these charts.



this reason, prior knowledge of the DGP is essential.

1.8 And now *Causal Diagrams* come

At this point we are clear about what we want to achieve and we are aware of the risks involved if we do not first acquire a solid knowledge of the DGP. What we have left to do is to set the research design. How to do this? How to figure out whether to resort to a *Randomized Controlled Trial* or to an *Observational Method*⁴? And are we sure that we will be able to complete the experiment by adopting that set of actions/operations?

The point is, therefore, to understand which econometric tools to use or, more generally, what actions we need to take to obtain identification, before to start "digging in the data" [Hun21]. A bit like building a house: nobody builds a house without the architect's project. The answer to this question - "How to set the research design?" - depends on how the DGP is structured.

So, the following question is: "How do we get a clear idea of how the DGP is structured?" Easy: *causal diagrams*! Okay, maybe not so easy, but causal diagrams are a particularly effective tool to set up research design, and not only that.

Causal diagrams are graphs that allow us to immediately visualize how the DGP is structured. Here are an example of a complex causal diagram. Imagine you have to keep in mind a DGP like this. Impossible task.

⁴These are some Econometrics type of tools.

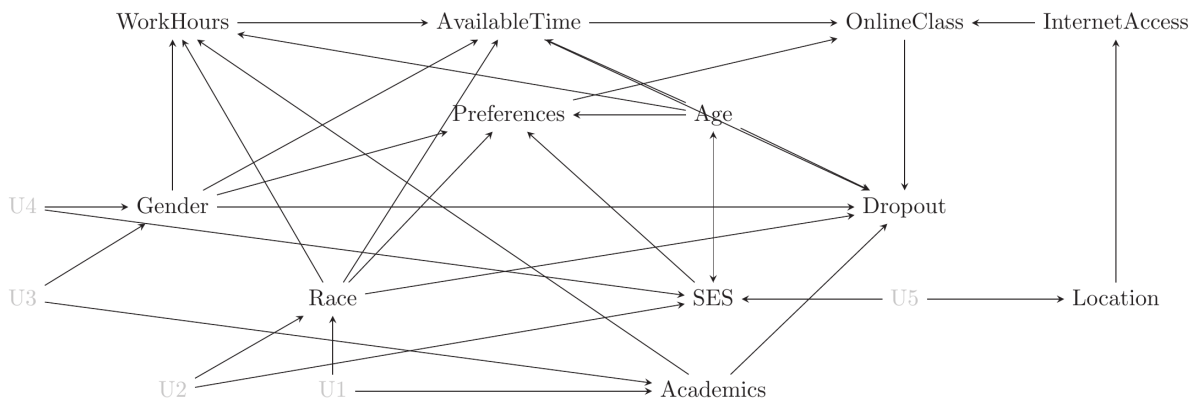


Figure 1.1: Huntington-Klein [Hun21], figure 7.1, p.106

Once we understand how the DGP is structured we also have sufficient information to decide what statistical-econometric tools to apply [Hun21, p.87, 115]. This is also important from data collection point of view: collecting data is in fact a costly operation in terms of time and money, so it is appropriate, before the implementation of research design, to have clear whether and which data to collect, so as to avoid waste of time and money. Moreover, it must be considered it's not always possible to obtain identification. This is generally due to:

The impossibility of collecting data because a certain variable is not measurable.

For example, the individual's ability is definitely a relevant factor in choosing whether to continue in the course of study, but how can we uniquely measure the individual's ability?

The impossibility to collect data because it would be too expensive.

To understand whether a glass of wine a day is good or bad for our body, we should ask all the subjects involved in the experiment what they eat and drink in addition to the glass of wine. It would have an excessive cost to ask thousands of individuals to record what and how much they eat and at what time of the day [See Hun21, chapter 5.4].

Therefore, causal diagrams allow us to understand, before starting the research, how we think to obtain identification or if we cannot answer the causal research question with the tools available.

So we figured out what causal diagrams are for, but how do we build them? How are they used specifically? We will talk about this in the following chapters.

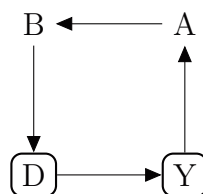
Chapter 2

Causal Diagrams: how are they made?

2.1 Causal Diagrams or *Directed Acyclic Graphs*?

The causal diagrams we are going to analyze now are more properly called *Directed Acyclic Graphs* (then shortened **DAGs**). "They were developed in the mid-1990s by the computer scientist Judea Pearl [Pea09] who was trying to develop a way for artificial intelligence to think about causality" [Hun21, p.90].

DAGs are so called because they cannot contain cyclic causal effects. Here an example of a cyclic causal effect.



2.2 Basic elements: *Nodes* and *Arrows*

Causal diagrams are composed of two elements: nodes and arrows.

Nodes represent variables within the DGP. Each variable can take multiple values [Hun21, p.91], but the way it will always be drawn is the same: a node. I'm going to draw Determinant(D) and Outcome(Y) variables within a rounded rectangle, in order to recognize them quickly.



All relevant variables in the DGP should be included, although we cannot measure or see them. This pops up all the time in social science [Hun21, p. 93]. The variables that cannot be measured are called *unobserved variables* or *unmeasured variables*. I'm going to indicate them and the arrows that comes out from them with a light gray color.

$$A \rightarrow \text{Unmeasured} \rightarrow B$$

Arrows show the causal relationships between variables, between nodes. They only tell us that one variable causes another, but nothing says about the sign and the type¹ of the causal effect [Hun21, p. 91].

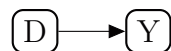
Potentially everything could affect everything, but we need to focus just on the most relevant relations [Hun21, p. 91]. Importance is subjective, and so DAGs could differ just because two scientists have a different opinion about the importance of some variables of the same DGP. *Along the lines of “focusing on important stuff”, causal diagrams are often drawn with a particular outcome variable in mind. This is done because it allows you to ignore anything that is caused by that outcome variable*² [Hun21, p. 91].

Another important thing to underline is that, when one variable is caused by multiple things, the diagram does not tell us exactly how those things come together³ [Hun21, p. 93].

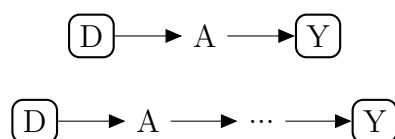
2.3 Causal Effects

The causal effects that can be shown in DAGs are of two types and cannot be cyclic:

Direct Causal Effect: when simply a variable affects another variable [Hun21, p. 96].



Indirect Causal Effect: when the causal effect between two variables is mediated by a third variable (or more variables) [Hun21, p. 98]. Variables that interpose between the Treatment (D) and the Outcome (Y) are called *Mediators*.



¹It could be linear, quadratic, exponential and so on.

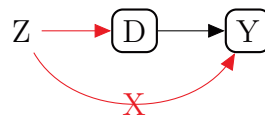
²This point is explained better in [chapter 3](#).

³This aspect becomes so relevant when thinking about *moderators*.

2.4 Fundamental assumptions

When there is an arrow this means that between those two variables there is a direct causal relationship, but when the arrow is not there? “No arrow means no *direct* causal effect” [Hun21, p.96]. This is a fundamental assumption.

As we said, there is a balancing act between omit or not [Hun21, p.96, 97] which is essentially subjective. But, there are some assumptions on which is based the possibility of obtaining identification, i.e. to answer the causal research question. These assumptions, which may consist of the presence or absence of certain arrows between nodes, are called *identifying assumptions*. These are important because, if they are incorrect, the approach adopted to identify the answer of our research question is not going to work [Hun21, p.97]. For example, to run an *Instrumental Variables* research design you need the *instrument* (Z) to be *relevant* (the $[Z \rightarrow D]$ must be different from zero) and *exogenous* (the *exclusion restriction* $[Z \rightarrow Y]$ must be zero). If one of these two is not satisfied, we cannot use the IV econometric tool. The two conditions are *identifying assumptions*.

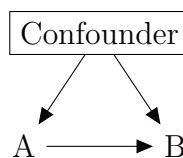


2.5 Common structures

Normally some particular structures turn up in DAGs. Recognize them is fundamental to realize a research design that allows us to obtain identification.

2.5.1 Confounders

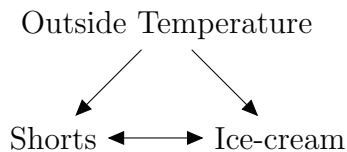
In statistics, a **confounder** is a variable that influences both the dependent variable and independent variable, causing a spurious association [Wikc]. So, when we talk about confounding structures with reference to DAG, the typical structure is this one⁴:



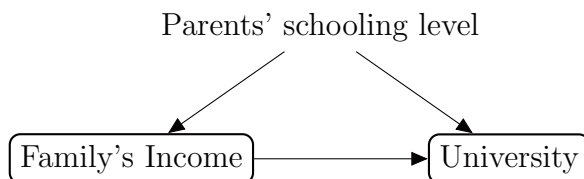
Confounders are one of the reasons why *correlation does not imply causation*. For example, *eating ice-cream* and *wearing shorts* are two positively related variables, but none of us would dare to say that one is the cause of the other. If we did, we would be wrong,

⁴Note that the confounder is a structure that do not necessarily affects the D and Y. It is simple a structure between three variables.

because it is the *outside temperature* that determines whether to eat ice cream or whether to wear shorts. So the outside temperature is the *confounding variable* [Hun21, p. 94].



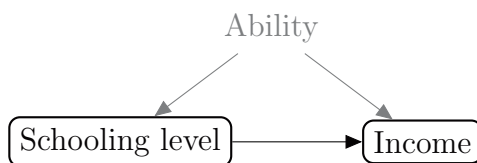
Let's come back to our [problem](#) of identifying the causal effect of *family's income* on the *individual's probability of enrolling in academic studies*. *Parent's schooling level* is a confounder. Why?



On one hand, we can assume that parents' schooling level affects the family's income because we can assume, on average, the higher the parents' schooling level, the higher the family's income.

On the other hand, we can assume that parents' schooling level affects the child's probability of enrolling in academic studies. Why? Because of the so-called *family background*: the higher the level of education of the parents, the greater the incentives of these for the children to study at least as much as they did.

Talking about confounders, it's important to underline that often confounders are also unmeasurable variables. For example, *individual's ability* is an unmeasurable variable and it's reasonable to assume it affects both the *level of education acquired* and the *income earned* by the individual during working age.

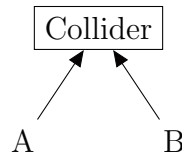


Variables such as *individual's ability* that represent general concepts and that cannot be measured are called **latent variables** [Hun21, p. 94].

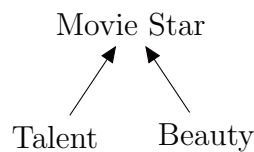
2.5.2 Colliders

In statistics, a variable is a **collider** when it is causally influenced by two or more variables. The name "collider" reflects the fact that in graphical models, the arrow heads

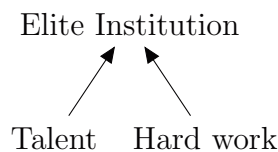
from variables that lead into the collider appear to "collide" on the node that is the collider [Wikb].



One example to understand colliders is about movie stars. To be a *movie star*, you should be either talented or beautiful, or both. So, *movie star* is a *collider* [Cun21, Chapter 3.1.6].



Another example can be given by whether or not you are part of an *elite institution*. To enter a high-level institution you are either talented or you have worked hard. So, being in an elite institution is a collider [Aga].



Notice that, generally, a variable is a collider when the following logical reasoning occurs: "The collider variable is caused by either the value of A (cause 1), or the value of B (cause 2), or both".

We will see [later](#) colliders have some properties for which attention should be paid.

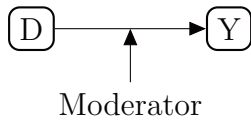
2.5.3 Moderators

Moderators are variables that modify the effect of one variable to another one [Hun21, p. 99].

Moderators vs Mediators

We must keep in mind the difference between *moderator* and *mediator* concepts: *moderators moderate* the effect that one variable cause to another; *mediators* are variables that *explain* how one variable cause another, because they interpose between the causing variable and the caused one.

Moderator variable



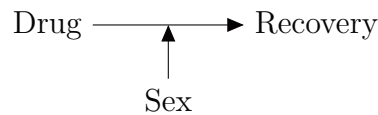
Mediator variable



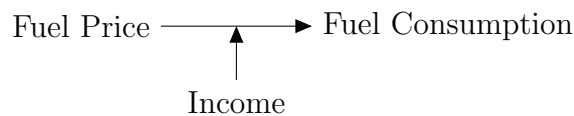
The two concepts are not mutually exclusive. Though, it's important to recognize when a variable is a moderator and/or a mediator, because this affect how we will manage that variable.

Moderators examples

For instance, we want to calculate the effect the *administration of a certain drug* has on the *recovery from a pathology of the uterus*. Clearly, the effect will be moderated by the *sex* variable: only on those patients who have the uterus the effect will be observed.



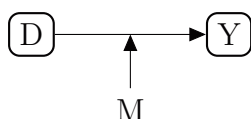
The moderators are not always so clear. Let's think about the effect of an increase in the cost per litre of fuel on fuel consumption. This effect is moderated by income: those with a higher income will suffer less from the upward changes in the price of fuel.



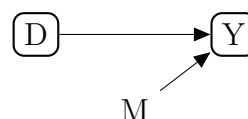
Problematic notation

Probably you have noticed that in the previous graphs the moderator is indicated with an arrow that affects another arrow. This notation, though intuitive, is however incorrect. Moderators in DAGs should be drawn as all other variables.

Not correct



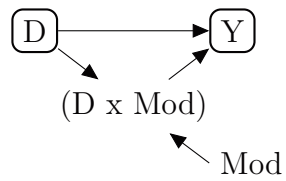
Correct



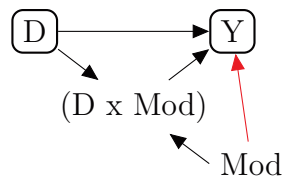
So, we have the problem of recognizing when a variable is a moderator, because the correct notation does not imply the variable to be a moderator. It could be simply a variable that causes another variable (in this case the outcome).

A trick that can be adopted, as Huntington-Klein [Hun21, p. 100] suggests, is to insert

the *moderated effect* ($D \times \text{Mod}$) right on the diagram, as a *mediator*. The solution is not formally correct, but it makes the interpretation of the DAG much more intuitive.



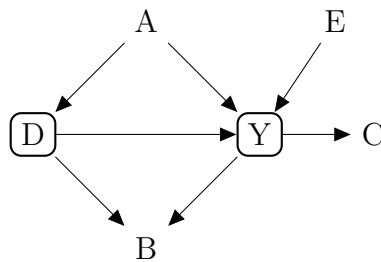
In the previous DAG, M , the *moderator*, has only a moderating effect on D . It's not to exclude, however, that a moderator can have its own causal effect - in addition to the moderating one - on other variables and, why not, also on the outcome (see the red arrow).



Chapter 3

How to draw Causal Diagrams?

Before starting to draw a DAG, it is essential to study the institutional environment of our causal research question, that is to understand as much as possible about the variables and the causal relationship of our DGP. Once we have done this or, at least, when we think we have reached a good comprehension of our DGP, we can start drawing it. In the first place we insert the treatment and the outcome variable. Then, we add all the relevant variables in the DGP and finally we trace all the causal relationships between the variables. Remember that we are focusing on the causal effect of treatment on outcome, so every variable that is caused by the outcome and it is not connected in some way with the treatment should be removed. In this example, the variable C should be removed.



In like manner for variables like E, that causes (and not caused by) the outcome and are not connected to every other variable. This kind of variables should be removed too, because they are not relevant to our objective of finding the effect of D on Y. However, even if variables like C and E should be removed, initially it could be convenient to keep them, because we could find a relevant relationship later.

Drawing suggestions I have reported here come from “The Effect: An Introduction to Research Design and Causality” [Hun21, chapter 7], where you can find a more detailed guideline to draw causal diagrams.

Chapter 4

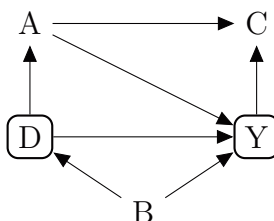
How to use Causal Diagrams?

4.1 Paths: what are they?

We have understood DAGs answer the purpose to set our research designs, but how precisely? Studying *paths*.

Paths are the possible ways that, in a DAG, link the treatment variable (D) to the outcome variable (Y) [Hun21, p. 116]. Graphically, paths are all possible *roads* that start from the treatment and arrive at the outcome. For example, in the DAG below, all the possible paths are:

- I $D \rightarrow Y$
- II $D \rightarrow A \rightarrow Y$
- III $D \rightarrow A \rightarrow C \rightarrow Y$
- IV $D \leftarrow B \rightarrow Y$ ¹



So, each path shows us a possible way for which treatment and outcome are connected to each others, but not necessarily by a causal link². In order to obtain identification, we will decide which paths we are interested in and which are not and, consequently, we will set our research design [Hun21, p.116].

¹Note that paths, to be such, do not necessarily have to have arrows all pointing in the same direction.

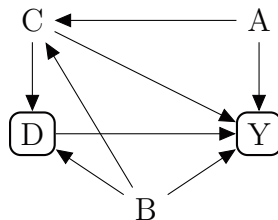
²For example, look at path IV where B is a confounder.

4.2 How to find all the paths

In the previous example, recognizing all paths was relatively simple. But when DAGs become complex, how can we be sure that we haven't forgotten any of them? We use a guideline that allows us to detect all of them at first glance. Nick Huntington-Klein, in his book "The Effect", has developed the following algorithm [Hun21, p.117]:

1. *Start at the treatment variable (D)*
2. *Follow one of the arrows coming in or out (either is fine!) of the treatment variable to find another variable*
3. *Then, follow one of the arrows coming in or out of that variable*
4. *Keep repeating step 3 until you either come to a variable you've already visited (So there is a loop! It cannot be in DAGs), or find the outcome variable (that's a path. Write it down)*
5. *Every time you either find a path or a loop, back up one and try a different arrow in/out until you've tried them all. Then, back up again and try all those arrows*
6. *Once you've tried all the ways out of the treatment variable and all the eventual paths, you've got all the paths!*

This steps sequence allows us to recognise each path in the DAG we are working on. Let's try to apply it to the following DAG.



Let's start from D and arrive at Y. This is the first path: $[D \rightarrow Y]$

Let's return to D and move towards C. From C we go to Y. This is the second path: $[D \leftarrow C \rightarrow Y]$.

Given that C is linked to other nodes, let's return to C and from here we move towards A. A has a single link to Y, hence the third path is: $[D \leftarrow C \leftarrow A \rightarrow Y]$.

In A there are no links to other variables other than C or Y, so we return to C. C also connects with B. We notice B in turn goes to Y. The fourth path is: $[D \leftarrow C \leftarrow B \rightarrow Y]$.

But from B another link goes to D. Then, the path would be $[D \leftarrow C \leftarrow B \rightarrow D]$

that is a loop! Therefore it's not a path, because in DAGs there cannot be cyclic paths. So we have to ask ourselves if the loop can be eliminated. If it cannot be suppressed because the effect that the loop describes is relevant for our research, then we have to adopt other tools that are not the Directed Acyclic Graphs, because in these there cannot be loops.

In B the links to other variables are finished, so we return to C. In C also the connections to other variables are finished. Let's return to D. From D there is a link that we hadn't considered so far: $[D \leftarrow B]$. Then, from B you can connect to Y. The fifth and final path is: $[D \leftarrow B \rightarrow Y]$.

At this point we have identified all the paths present in our DAG:

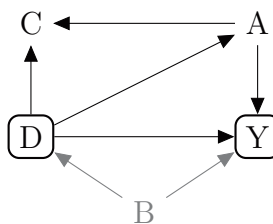
- $D \rightarrow Y$
- $D \leftarrow C \rightarrow Y$
- $D \leftarrow C \leftarrow A \rightarrow Y$
- $D \leftarrow C \leftarrow B \rightarrow Y$
- $D \leftarrow C \leftarrow B \rightarrow D$ (loop)
- $D \leftarrow B \rightarrow Y$

4.3 Categorize the paths found

Once all paths are recognised and all cycles are eliminated - assuming they can be eliminated because they're not sufficiently relevant for our research - we need to categorise paths to understand how to treat them.

4.3.1 *Front door and Back door paths*

A first distinction that can be made is between front door and back door paths. **Front door paths** are the paths where all the arrows face away from the treatment [Hun21, p.121]. Conversely, **back door paths** are the paths where at least one arrow, somewhere along the path, points towards the treatment variable [Hun21, p.121] Consider the following DAG.



Front door paths are:

- $D \rightarrow Y$
- $D \rightarrow A \rightarrow Y$

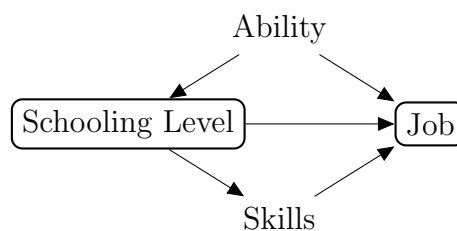
Whereas back door paths are:

- $D \rightarrow C \leftarrow A \rightarrow Y$ (note that C is a *collider*)
- $D \leftarrow B \rightarrow Y$ (note that B is a *confounder*)

Note that, when there is at least a *collider* or a *confounder* in the path, this will definitely be a *backdoor path*.

4.3.2 Good and Bad paths

Front vs back door paths subdivision is useful to give a name to these two types of paths, but since our goal is to obtain identification, and therefore to set the research design, the subdivision between good paths and bad paths becomes more important. A causal path is a **good path** if it is related to our research question, so, a path we are trying to identify [Hun21, p.121]. Conversely, a **bad path** is a path that does not count for our research question [Hun21, p.121]. So, it is the relevance with respect to our research question the determinant of the paths goodness. For example, let's consider this DAG where the *individual's schooling level* is the treatment and the *probability of being hired for a job* the outcome.



The front doors are two. What's their interpretation?

schooling level \rightarrow **job** : this is the so-called *signalling effect*, for which future employees seek to increase their probability to be hired by achieving higher educational degrees. It is based on the assumption that the higher the individual's schooling level the higher the likelihood of being hired [Wike]. Let's call this path the *signalling path* for convenience.

schooling level \rightarrow **skills** \rightarrow **job** : this path is telling us the probability of being hired depends on educational level, but because it increases the future employee's skills. Let's call this path the *skills path* for convenience.

Think of a firm that needs to hire a manager. It selects some candidates and then has to decide which one to hire. The candidates' probability of being hired depends on:

the educational level achieved : individuals with an executive degree have a greater chance of being taken on rather than others with only a postgraduate degree, regardless of their management ability³. This is the signalling path.

the management skills acquired during the studies : the higher the work ability acquired thanks to the studies, the higher the probability of being hired. This is the skills path.

The weight of the paths depends on how the job selections and interviews are conducted. If our research question is “How does the manager candidates' schooling level determine their probability of being hired?” therefore good paths are [schooling level → job] and [schooling level → skills → job]. On the contrary, if our research question is “How does the manager candidates' schooling level determine their probability of being hired *because of the signalling effect?*” hence, the only good path is [schooling level → job]. Instead, if our research question is “How does the manager candidates' schooling level determine their probability of being hired *because of the management tools acquired during their educational path?*” so, the only good path is [schooling level → skills → job].

To decide whether a path is good or not we need to clearly keep in mind what is our research question. In our managers example, depending on the causal research question, we will adopt different strategies to get identification.

4.4 Get identification: to close all the *bad* paths

We have just said there are good and bad paths. We are interested only in good paths, that is we want to identify only the causal effect described by these ones, excluding the bad paths. How to do this? The gold rule of identification says: “*To get identification you need to close all the bad paths and leave the good ones open*”. This definition is implicitly saying that there are *open* and *closed* paths. What does this mean? Let's see this distinction better.

4.4.1 *Open and Closed paths*

Imagine a path like a river, which starts to a certain point, the “treatment”, and flows into the “outcome”. Along the river are located the path's variables. The treatment variable produces a wave, that is the “causal effect” produced by the treatment itself. If

³They may have obtained the degree by copying every exam.

there are no obstacles, the wave advances through the variables, until it arrives at the “outcome” one, producing its causal effect. Vice versa, if there is, let’s say, a dam, the wave is blocked before to kick the outcome. If there are no obstacles the path is *open* and, conversely, if there are, the path is *closed*. To give a formal definition: “A *causal path is open* if all the variables along the path have variations in the data and a *causal path is closed* if at least one of the variables along the path has no variation in the data” [Hun21, p.122]. So, only if the path is open the causal effect can propagate along the path and finally impact into the outcome.

With this trivial example we have explained the concept of **d-separation** (or **d-connectedness**):

“If all the paths between two nodes X and Y are blocked, then we say that X and Y are d-separated. Similarly, if there exists at least one path between X and Y that is unblocked, then we say that X and Y are d-connected. [...] d-separation is such an important concept because it implies conditional independence.” [Nea20, p.29]

So, when the path is open there is a causal relationship between the treatment and the outcome and when the path is close, well, no.

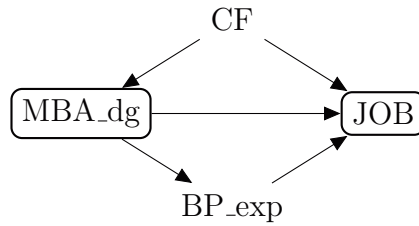
Now, maybe, you are glimpsing how the story ends up: blocking all the bad paths we will be able to observe only the causal effect we want to know. In other words, to close the bad paths is a way to skim each distortion in our data that does not allow us to see the searched causal effect.

Let’s return to our [previous example](#). If our research question is “How does the manager candidates’ schooling level determine their probability of being hired *because of the signalling effect?*” what we have to do, to get identification, is to close all the “roads” except [schooling level → job]. The next question is: “How to practically close a path?”

4.4.2 How to close a path? *Controlling for*

Controlling for (or **adjusting for**) a variable is a way to “remove all the variation associated with that variable (the **controlled variable**) from all the other variables” [Hun21, p.59]. *Controlling for* is a complex concept and our example can provide us an easy way to catch it.

Let’s take an example similar to the manager candidates one. MBA_dg is the *Master in Business and Administration degree grade*. BP_exp is the *expertise in making a business plan*. JOB is the *candidate’s probability of succeeding in application*, i.e. the probability of obtaining the job. CF are other confounder factors that we suppose to be measurable.



We want to know “How does the master degree grade affect the probability of being hired *because of the signalling effect?*” Our hypothesis is that firms, when searching for a manager, hire someone or not because he/she achieved a certain degree grade, regardless of his/her real ability in making a business plan⁴. This is the signalling effect.

To identify our answer, we select a sample of MBA graduate students and a sample of firms that are searching for a manager. Then, we invite each student to apply for the manager job position in every firm inside our sample. We collect data on candidates’ degree grade, i.e. *MBA_dg*, and their percentage value of job applications in which they succeeded in the interview, obtaining a job offer (*JOB*). *MBA_dg* takes values from 1 to 4, in relation to the degree grade class: 1 if “fail/borderline” (final grade of 40 - 49 %); 2 if “pass” (50 - 59 %); 3 if “merit” (60 - 69 %) and 4 if “distinction” (70+ %). Then, we *run a regression* like this:

$$JOB = \beta_0 + \beta_1 \cdot MBA_dg + \mathcal{E} \quad (4.1)$$

Suppose we get this result:

$$JOB = 0.21 + 0.15 \cdot MBA_dg + \hat{\mathcal{E}} \quad (4.2)$$

That 0.15 means the probability of being hired increases by 0.15 for each class upgrade. Did we get identification? No, because we are not *controlling for* *BP_exp* nor for *CF*. So the 0.15 is not the real causal effect value of the *signalling effect*. To get it we *adjust* our regression in order to *control for* *BP_exp* and *CF*. How? Collecting data on *BP_exp* and all the confounder factors and adding them to the regression as *control variables*. Imagine the *BP_exp* is measured with a test that is administered to all the candidates sample⁵ in order to have a uniform ability assessment. Suppose in the test students can score 3 possible values: 1 if “low expertise”; 2 if “average” and “3 if “high”. Then, we run the regression with new data and obtain this:

$$JOB = 0.02 + 0.04 \cdot MBA_dg + 0.07 \cdot BP_exp + 0.05 \cdot CF + \hat{\mathcal{E}} \quad (4.3)$$

⁴Even if the signalling effect exists, this is not a realistic hypothesis, but we are taking it for real just for the sake of providing a clear example.

⁵Suppose we have all the money we need to conduct this experiment.

Did we get identification? Yes, because we have closed all the bad paths controlling for at least one variable in these. And so, the answer to our research question is that the signalling effect is 0.04 on the MBA_dg class upgrade. So, returning to our question “What is *controlling for*?”, in this example controlling for BP_exp we removed the source of variation produced by MBA_dg that impacts on JOB through BP_exp and controlling for CF we also skimmed the influence caused by the other confounder factors (CF). This is also visible looking at the data: the MBA_dg coefficient decreased from 0.15 to 0.04. Why? Because adding other control variables the regression “was able to assign more precisely to each variable the amount of variation it produced”. In other words, adding the control variables regression was able to distinguish the causal effect produced by each variable, instead of assigning only to MBA_dg the most of the variation that the outcome, JOB, showed in the data collected. As a matter of fact, the 0.04 coefficient has to be read as: “The MBA_dg increases the probability of being hired by 0.04 *ceteris paribus*, that is keeping constant all the other variables”.

Adding a variable in a multiple regression is the most common way to *control for* that variable. However, there are other methods that allow us to control for a variable. Which one to choose depends on your causal question.

4.4.3 Pay attention to *Colliders*

When we close paths to get identification we need to pay attention to colliders because of two reasons.

- They close the path where they are by default;
- If we control for them, we open up the paths.

Path closed by default

You may be asking why this happens. Nick Huntington-Klein provided us this intuitive explanation:

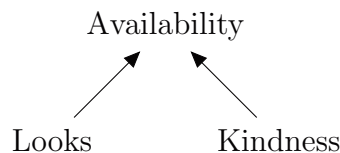
“You can think of it this way: the collider variable doesn’t cause anything else on the path. It’s just being caused by the variables to its left and right on the path. So if we’re looking for alternate explanations of why Treatment and Outcome might be related, the collider shuts that alternate explanation down. If the path were [Treatment ← C → Outcome], without a collider, then one reason why Treatment and Outcome vary together is because C causes them both. But with a collider, [Treatment ← A → B ← C → Outcome], C can affect Outcome, and C can affect B, but because B doesn’t affect Treatment, C can no longer induce a relationship between

Treatment and Outcome. B saved us. [Hun21, p.124, 125]

Controlling for colliders open up paths

Controlling for a collider opens up the path/s *because once you control for the collider the two variables pointing to the collider become related*⁶ [Hun21, p.125]. Brady Neal gives us an intuitive example to explain this collider behavior:

“Imagine that you’re out dating men, and you notice that most of the nice men you meet are not very good-looking, and most of the good-looking men you meet are jerks. It seems that you have to choose between looks and kindness. In other words, it seems like kindness and looks are negatively associated. However, what if I also told you that there is an important third variable here: availability (whether men are already in a relationship or not)? And what if I told you that a man’s availability is largely determined by their looks and kindness; if they are both good-looking and kind, then they are in a relationship. The available men are the remaining ones, the ones who are either not good-looking or not kind. You see an association between looks and kindness because you’ve conditioned on a collider (availability). You’re only looking at men who are not in a relationship.” [Nea20, p.26, 27]



And here is the graphical representation of his example, where he assume “looks” and “kindness” independent:

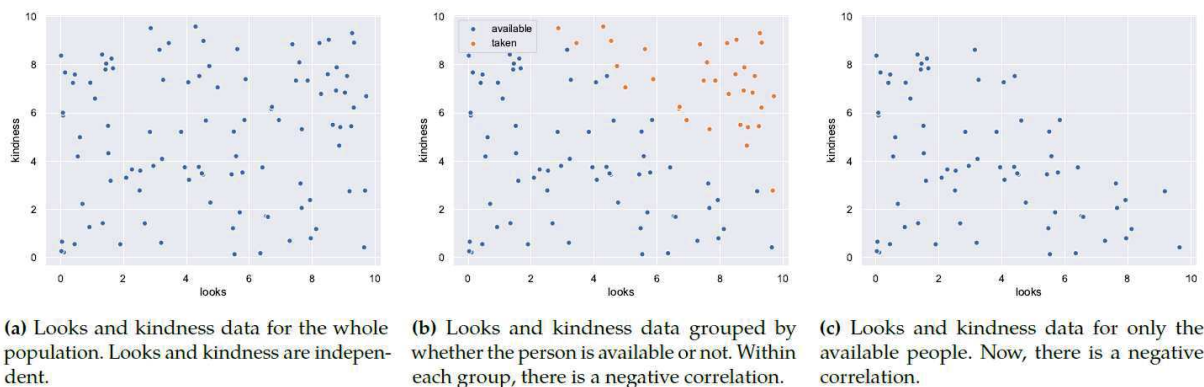
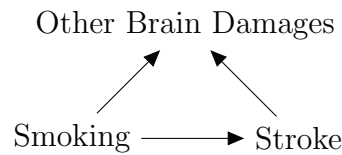


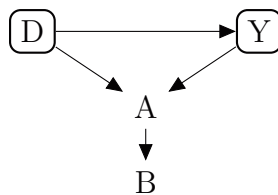
Figure 4.1: [Nea20, Figure 3.18, p.27]

⁶The formal explanation of this collider behavior is called “Berkson’s Paradox” [Wika].

Even if it seems strange, it is misleading to control for every variable, because we could accidentally control for a collider. For instance, we want to estimate how smoking affects the probability of having a stroke.

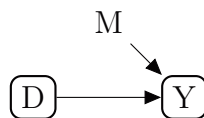


If we regress stroke data on smoking data we can get identification according to this DAG. But, if we also control for having other brain damages, then the path [smoking \rightarrow other brain damages \rightarrow stroke] opens and then we do not have identification. *Conditioning on descendants of a collider also induces association in between the parents of the collider.* [...] *In other words, a descendant of a collider can be thought of as a proxy for that collider, so conditioning on the descendant is similar to conditioning on the collider itself.* [Nea20, p.28] In the following DAG, for instance, if we control for B, then [D \rightarrow A \leftarrow Y] opens up.



4.4.4 Pay attention to *Moderators*

When controlling for variables we need to pay attention to *moderators* because the relationship between a moderator and its *moderated variable* can take multiple forms, since the moderator could describe different types of relationship⁷. For example [Hun21, p.99], consider the following DAG, where M is the moderator⁸:



If we translate this DAG in a regression, it could take different forms because we do not know how the moderator is affecting other variables:

- I $Y = \beta_0 + \beta_1 D + \beta_2 M + \mathcal{E}$
- II $Y = \beta_0 + \beta_1 D + \beta_2 M + \beta_3 M^2 + \mathcal{E}$
- III $Y = \beta_0 + \beta_1 D + \beta_2 M + \beta_3 MD + \mathcal{E}$
- IV $Y = \beta_0 + \beta_1 D + \beta_2 MD + \mathcal{E}$

⁷Check on [Problematic Notation](#).

⁸We're now using the right notation for moderators.

Note that only in III and IV there is a moderating effect⁹. So, which is the right regression? It depends on our interpretation of the DGP. As you probably are realizing, to get identification we need to take the *right* regression. This is the reason why we need to pay attention to *moderators*.

4.4.5 The *Backdoor Criterion*

In DAGs theory it is more common to hear about “backdoor paths” instead of “bad paths”, because usually the bad paths coincide with the backdoor paths. Given that, the gold rule to get identification is called ***backdoor criterion***, and it says:

“A set of variables W , satisfies the backdoor criterion relative to D and Y if the following are true:

1. *W blocks all backdoor paths from D to Y .*
2. *W does not contain any descendants of D ¹⁰ [that are situated in front door paths¹¹] [Nea20, p.37]*

If W satisfies the backdoor criterion we say that W is a sufficient adjustment set because W is sufficient to adjust for to get the causal effect of D on Y [Inf20].

4.4.6 To Sum Up

Here’s a recap about how to block a path, taken from “Introduction to Causal Inference” of Brady Neal [Nea20, p.28]:

“A path between nodes D and Y is blocked if either of the following is true:

1. *Along the path there is a chain $[... \rightarrow W \rightarrow ...]$ or a fork $[... \rightarrow W \leftarrow ...]$ where W is conditioned on.*
2. *There is a collider W on the path that is not conditioned on and none of its descendants are conditioned on.*

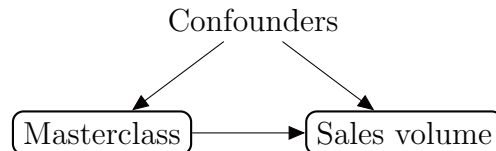
⁹Where there is the MD explanatory variable.

¹⁰because in the Backdoor Criterion all the front door paths are considered good paths, but we have seen in [MBA_dg \rightarrow JOB] example that is not necessarily true.

¹¹The part in squared brackets is added by me.

4.5 The *Placebo Test*

So far we have talked about how to use DAGs to get identification, supposing the DAGs we draw are complete, but it might happen we forget to put some *relevant* causal relationships in our DAGs. How to be sure that we are not forgetting anything relevant? ***Placebo Test***. How does it work? An example is the best way to explain it. Consider the following graph, where we want to evaluate the effect of a *masterclass participation* on the *individual's sales volume* of a sales team.



To measure the causal effect we split the sales team members into two groups: a ***treatment group*** - that is a group made up of half of the sales team members and that will attend the masterclass (the treatment) - and a ***control group*** - that is a group made up of the other half of the sales team members and that will not receive the treatment. What are we going to do is to compare, after the masterclass attendance (so after the treatment is administered), if the sales volume (outcome), on average, is higher for the treatment group compared to the control one. We can compare the two values of sales volume only if the two groups are similar except for the masterclass attendance or not. So, to *control for* every possible difference we adopt some strategy in order to keep track of every possible confounding variable. In practice, we control for confounders into our regression.

$$Sales_volume = \beta_0 + \beta_1 Masterclass + \beta_2 Confounders + \mathcal{E} \quad (4.4)$$

How to be sure we're controlling for every relevant confounder? We carry out a *Placebo Test*. How? We run the previous regression using *Sales_volume* measured *before* the *Masterclass attendance* as a dependent variable and check the value of β_1 . If we have controlled for confounders in the right way, we'll see $\beta_1 = 0$, which means there are no factors, except for the treatment administration, that affect the sales volume. On the contrary, if we'll see $\beta_1 \neq 0$, this means we are not controlling for some relevant variables in confounders, and so we need to check which one is missing. This could happen, for instance, if we are not able to accurately control for *engagement*, which reasonably affects the masterclass effectiveness and the individual's sales volume.

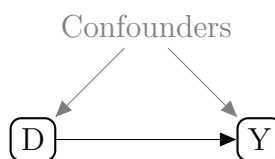
Chapter 5

Not only closing backdoors

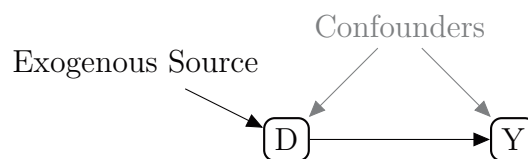
You can think at this point DAGs are an irreplaceable tool for causal inference. That's partially true and depends on how we are going to use DAGs. As we have seen so far, they are a useful tool to identify backdoor paths and close them. However, not always the DGP we are dealing with allows us to apply the *backdoor criterion*, that is closing all the bad paths to get identification. This could be due to unmeasured variables that block us in every possible way to close all the bad paths. So, in this type of situation, DAGs are useless tools? For closing backdoor paths yes, but they can help us in finding another *strategy* to get identification.

5.1 Exogenous source of variation

Suppose you are dealing with a DAG like this.



There is no hope of closing backdoor paths. So, how can we get identification? One way to get identification is to recognize an *exogenous source of variation that determines only D*. What does this mean? Graphically this:

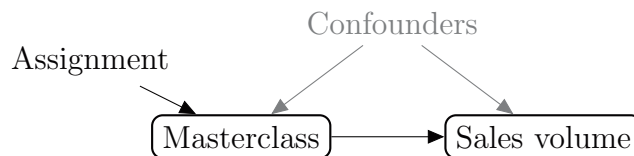


So, it is a variable that triggers only the D, without being affected by other variables¹.

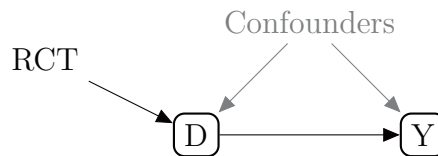
¹“Exogenous” means “without being affected by other variables” and “without affecting variables”

Why is this variable useful? Because it allows us, in some way, to close all the backdoor paths and obtain the causal effect $D \rightarrow Y$. The *third variable* can be generated by the scientist or simply taken from real-world situations.

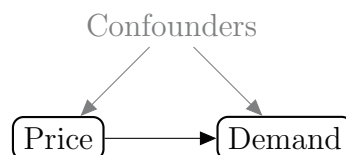
We talk about a *third variable generated by the scientists* when they assign or not some individuals to a certain treatment in order to create a *treatment group* and a *control group*. The [previous example](#) on masterclass and sales volume is an example of this kind of *third variables*, because, in order to evaluate if the business masterclass increases individual's sales, scientists assign *randomly* individuals of the sample (Sales team) to the masterclass or not and then compare their sales volume.



When scientists decide whether or not individuals of the sample group are randomly assigned to the treatment group or to the control one, we say this kind of experiment is a ***Randomised Controlled Trial (RCT)***.



On the contrary, we talk about *third variables taken by real-world situations* when an exogenous source of variation even if it is not produced by scientists but, likewise RCT, it allows us to answer our causal research question. This kind of *third variables* is called ***natural experiment*** because *randomization of the treatment occurs without a researcher controlling the randomization* [Hun21, p.133]. For example² [example inspired by this paper: AGI00], we want to estimate the *elasticity of tuna demand curve*, that is the percentage decrease in tuna demand³ because of a 1% increase in tuna price. We have data on the daily amount of tuna sold and the average selling price. The DAG is:

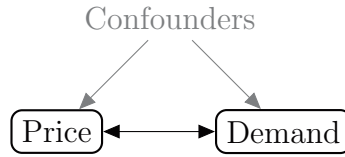


The problem is: “Are we sure that demand does not affect the price too?” Unfortunately except D”.

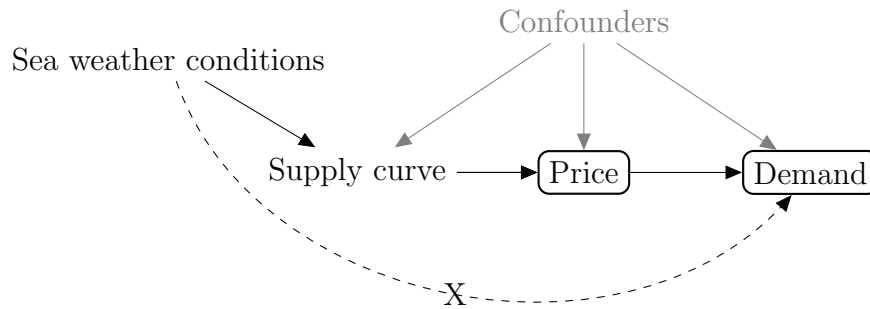
²In this example you won't recognise a control and a treatment. This could happen when working with treatments that are continuous variables, i.e. not binary as in the previous “masterclass - sales volume” example.

³Quantity demanded for consumption

not, because we know that in a competitive market quantity sold and price at which it is sold are simultaneously determined by the equilibrium between demand and supply. The DAG became so⁴:



How can we get out of this situation? We can resort to a natural exogenous source of variation: *sea weather conditions*. Sea weather conditions determine the availability of tuna, so the *tuna supply curve*, because the better the weather conditions the larger the quantity of tuna that can be sold. Then, given the basic economic demand fundamental for which “price is inversely related to quantity available”, we can figure out the *simultaneity* problem by tracing a new relationship that is [Sea weather conditions (Z) → Supply Curve → Price (D) → Demand (Y)].



With this research design we can identify the causal effect of price on demand, and, with some tricks, obtain the elasticity of the tuna demand curve⁵. We call *sea weather condition* the **instrumental variable** (usually marked with **Z**) and we say this kind of research design is an **Instrumental Variables Research Design (IV)**. To adopt an IV strategy we need the instruments to be *exogenous* and *relevant*⁶.

In conclusion, these examples aim to show you DAGs can be useful tools even if they cannot be used to close all the backdoor paths, i.e. resort to the *backdoor criterion*. DAGs don't help us running an RCT or IV, but they make it easier to figure out when to

⁴This type of problem where also the outcome affects the treatment is called **simultaneity**, and usually with instrumental variables research design the problem can be overcome.

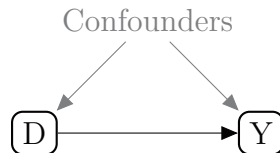
⁵Probably you are asking yourself why we cannot simply regress price data on demand ones and then measure the regression line slope, since the slope of the demand curve is its elasticity. It is not correct because data on price and demand are given by the intersection between the *demand curve* and the *supply curve* and we don't see if and how the supply and the demand curves *shift*. So, the sea weather condition is a variable that, for assumptions, causes only the supply curve shifting. This allows us to identify the points along the stationary tuna demand curve (because it's not influenced by sea weather conditions) and estimate the elasticity of demand of tuna. It's a tricky scenario, I know...

⁶We mentioned them exemplifying the [identifying assumptions](#).

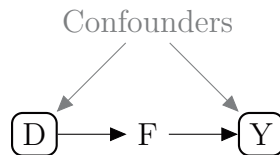
resort to RCT or IV and which *instrument* to consider. So, as I said in the first chapter, they help us set up the research design.

5.2 The *Front Door Method*

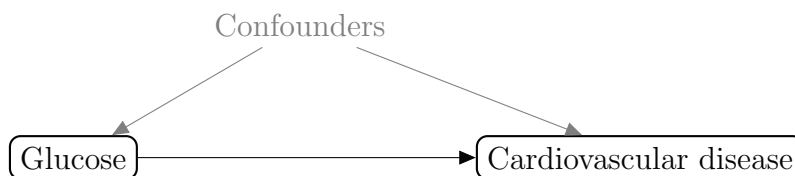
We have just seen DAGs help us understand where to adopt an RCT or IV research design. However, sometimes we cannot resort to these methods because there are no valid instruments or the RCT is not feasible. Hence, we are stuck with a DAG like this.



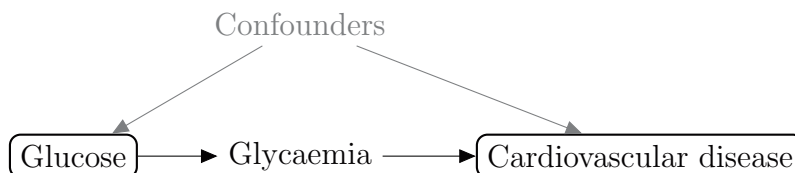
What can we do? The *Front Door Method* (FDM) is a solution we can pursue where DAGs like this:



The FDM tells us we can identify the causal effect $D \rightarrow Y$ by isolating separately $D \rightarrow F$ and $F \rightarrow Y$. For example⁷, we want to estimate the *glucose absorption* impact on the *individual's probability of contracting a cardiovascular disease*.



There are some confounders, such as *individual's stress*, that block us in closing backdoor paths. However, we know that glucose absorption increases *glycaemia level*, which in turn affects the probability of contracting a cardiovascular disease.



Adopting the FDM we can isolate the causal effects [Glucose \rightarrow Glycaemia] and [Glycaemia \rightarrow Cardiovascular disease] and therefore answer to our question [Glucose \rightarrow Car-

⁷The following example is inspired from another very similar in “The Effect” of Huntington-Klein [Hun21, p. 141 8.5].

diovascular disease].

The FDM works when the “variable in the middle” *has no link with other variables and captures a large portion of the reason why Treatment affects Outcome* [Hun21, p.141].

The FDM is another reason why DAGs could be useful even if we cannot use them to close all the backdoor paths.

Chapter 6

Conclusion

6.1 To Sum Up

As we have seen in this quick introduction guide, DAGs could be helpful in setting the research design and, if our strategy consists in closing all the backdoor paths, they are an effective starting point.

Now you are a step closer to winning the Economics Nobel Prize.

6.2 In short, my experience

During 2021 I attended an Econometric course held by professors Enrico Rettore¹ and Guglielmo Weber². During exam preparation, as every student does (I think), I tried to help friends and not only with some concepts that were a bit tricky. Unexpectedly, a *significant percentage* (as a statistician would say) of these students find my explanations really helpful. These kinds of situations happened in other exam preparations too, but this time my support to other people was much more significant. So, after the exam, I've started thinking about a master's degree in this field. Talking about this with Professor Rettore, he offered me the opportunity to do an internship at "[FBK-IRVAPP: Institute for the Evaluation of Public Policies](#)". There, with the help of Sergiu Constantin Burlacu³, I learned the fundamental concepts of Causal Inference⁴ by studying "Causal Inference: The Mixtape" [chapters 1 to 4] [Cun21] and "The Effect: An Introduction to Research Design and Causality" [Chapters 1 to 11] [Hun21]. Besides that, I supported the "[S.m.a.i.l.e. - Simple methods for artificial intelligence](#)" team in reporting activity

¹Rettore's [Scholar](#) page.

²Weber's [Scholar](#) page.

³Burlacu's [Scholar](#) page.

⁴RCT and the *observational methods*, that are IV, RDD, Diff in Diffs and Synthetic Control.

and I contributed to the drafting of a [proposal](#) for an impact evaluation on an artificial intelligence programme for financial market analysis. During these activities, I discovered DAGs and I found them a great potential tool to approach Causal Inference⁵. Then, I decided to write this Introduction to Causal Inference and DAGs remembering me helping my friends during exam preparation. I hope this work can help someone one day because yes, I have to say it: “I love helping people”.

⁵In fact, I wish I’d called this essay “Causal Inference for dummies”, but I thought this may get into some copyright issues.

Bibliography

- [Car99] David Card. “The causal effect of education on earnings”. In: *Handbook of labor economics* 3 (1999), pp. 1801–1863.
- [AGI00] Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. “The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish”. In: *The Review of Economic Studies* 67.3 (2000), pp. 499–527.
- [CK00] David Card and Alan B Krueger. “Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply”. In: *American Economic Review* 90.5 (2000), pp. 1397–1420.
- [Pea09] Judea Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009. DOI: [10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161).
- [Pin15] Paolo Pinotti. “The economic costs of organised crime: Evidence from Southern Italy”. In: *The Economic Journal* 125.586 (2015), F203–F232.
- [Pin17] Paolo Pinotti. “Clicking on heaven’s door: The effect of immigrant legalization on crime”. In: *American Economic Review* 107.1 (2017), pp. 138–68.
- [Inf20] Brady Neal - Causal Inference. *4.6 - The Backdoor Adjustment*. Youtube. 2020. URL: <https://www.youtube.com/watch?v=U1S8Rq8IcrY>.
- [Nea20] Brady Neal. “Introduction to causal inference from a machine learning perspective”. In: *Course Lecture Notes (draft)* (2020).
- [Cha21] United 4 Social Change. *Simpson’s Paradox: When Correlation Does Not Equal Correlation - Data - Graphs Series — Academy ...* Youtube. 2021. URL: <https://www.youtube.com/watch?v=ZgQ5oZnynSI>.
- [Cun21] Scott Cunningham. “Causal inference”. In: *Causal Inference*. Yale University Press, 2021.
- [Hun21] Nick Huntington-Klein. *The effect: An introduction to research design and causality*. Chapman and Hall/CRC, 2021.

- [Aga] Lovkush Agarwal. *Examples of collider bias*. [Online; accessed 08-July-2022] [last edit at the visualization: Feb 21, 2021]. URL: <https://lovkush-a.github.io/blog/data%5C%20science/causality/tutorial/2021/02/21/collider.html>.
- [Wika] Wikipedia. *Berkson's paradox*. [Online; accessed 27-July-2022] [last edit at the visualization: 26 April 2022, at 20:42 (UTC)]. URL: https://en.wikipedia.org/wiki/Berkson%5C%27s_paradox.
- [Wikb] Wikipedia. *Collider (statistics)*. [Online; accessed 08-July-2022] [last edit at the visualization: 22 October 2021, at 14:47 (UTC)]. URL: [https://en.wikipedia.org/wiki/Collider_\(statistics\)](https://en.wikipedia.org/wiki/Collider_(statistics)).
- [Wikc] Wikipedia. *Confounding*. [Online; accessed 08-July-2022] [last edit at the visualization: 22 June 2022, at 01:03 (UTC)]. URL: <https://en.wikipedia.org/wiki/Confounding>.
- [Wikd] Wikipedia. *Correlation does not imply causation*. [Online; accessed 06-July-2022] [last edit at the visualization: 1 July 2022, at 19:00 (UTC)]. URL: https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation.
- [Wike] Wikipedia. *Signalling (economics)*. [Online; accessed 27-July-2022] [last edit at the visualization: 4 June 2022, at 02:22 (UTC)]. URL: [https://en.wikipedia.org/wiki/Signalling_\(economics\)#Spence_1973:_%5C%22Job_Market_Signaling%5C%22_paper](https://en.wikipedia.org/wiki/Signalling_(economics)#Spence_1973:_%5C%22Job_Market_Signaling%5C%22_paper).