Università degli studi di Milano–Bicocca

Scuola di Economia e Statistica

Corso di Laurea in

Scienze Statistiche ed Economiche

Bayesian Causality

Relatore: Dott. Stefano Peluso

Tesi di laurea di:

Simone Maria Gervasoni

Matricola N. 880068

Anno Accademico 2024/2025

Contents

Ringraziamenti

Inserire qui gli eventuali ringraziamenti, altrimenti eliminare

Chapter 1

Introduzione

Il problema della causalità è di fondamentale interesse per capire meglio il mondo che ci circonda, perciò la domanda è stata affrontata da filosofi e scienziati. La sua definizione rigorosa nel campo statistico è stata solo affrontata molto recentemente, [proprio perché estremamente vago], da Neyman (1932) e poi sviluppata da Rubin negli anni '70, i quali hanno fatto riferimento al modello dei potential outcome spiegato nel capitolo **??**. Questo modello diverge concettualmente da alcune definizioni portate avanti da filosofi come John Stuart Mill che definisce la causalità come "the antecedent, or the concurrence of antecedents, on which [a given phenomenon] is invariably and unconditionally consequent", per Mill dunque possiamo dire che A causa B se e solo se ogni volta che succede A succede anche B . Questo modello di causalità è molto riduttivo e ignora la aleatorietà degli eventi, per questo dobbiamo introdurre il potential outcome model.

Chapter 2

DAGs

Introduciamo come primo argomento i DAGs o Directed Acyclic Graph, questi servono per schematizzare relazioni di causalità, che non verrà definità rigorosamente ora ma affrontata nel capitolo 3. Un DAG è un grafo formato da vertici che rappresentano le variabili in analisi e da archi che rappresentano la relazione di causalità le variabili. Viene detto aciclico perché partendo da un vertice e seguendo gli archi non si può tornare su quel vertice. Questi grafi non sono in grado di descrivere causalità reciproca simultanea $A \leftrightarrow B$ o feedback loop $A \rightarrow B \rightarrow A$ a meno che si inserisca un ulteriore vertice con lo stesso nome, anche se in questi casi non si consiglia di utilizzare questo tipo di grafici (Cunningham, 2021).

$A \longleftarrow B$

(a) Non valid DAG

$A \longleftarrow B$

(b) Non valid DAG

$A \longleftarrow B \longrightarrow C$

(c) Valid DAG

(I grafi sono da rifare)

Bisogna inoltre dire che i DAGs, essendo dei grafi sono di natura più sintetici, contengono molte informazioni non solo grazie agli archi disegnati ma anche grazie a quelli **non** disegnati, ad esempio il DAG **??** implica che $A \perp\!\!\!\perp C$ visto che non esiste un arco che li colleghi.

2.1 Vertici e Archi

Fare breve cappello

2.1.1 Vertici

Poniamo la C come la variabile *outcome* cioè la variabile di interesse per l'indagine (la salute dopo un determinato intervento). Poniamo A come la variabile *exposure* cioè quella di cui vogliamo quantificare l'effetto causale sulla variabile *outcome* (per esempio che tipo di intervento viene proposto). Analizziamo quindi un singolo vertice in questo caso B, gli unici modi in cui si può collegare ad A e C sono i seguenti:

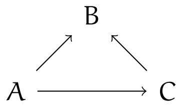

Figure 2.2: DAG: Common effect

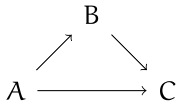

Figure 2.3: DAG: Mediator

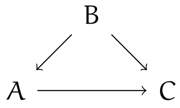

Figure 2.4: DAG: Common cause

Dunque B nel grafico 2.2 viene chiamato *collider*, mentre nei grafici 2.3 e 2.4 viene definito *non collider*, mentre solo 2.4 come *confounder*. Infatti se volessimo quantificare la relazione causale tra A e C quando B è un *non collider* o un *confounder* risulta più difficoltoso perchè sappiamo che B introduce sistematicamente correlazioni spurie tra A e C (bisogna capire meglio come affrontare i mediators [bisogna controllarli?], osservazione più giusta per i). Invece nel caso mostrato in figura 2.2 potremmo interpretare il $\hat{\delta}$ della regressione lineare:

$$C_i = A_i\delta + \epsilon_i \tag{2.1}$$

come quantificazione dell'effetto causale che A ha su C, questo lo possiamo affermare solo dopo aver confermato che effettivamente:

- il grafo soddisfa *Backdoor criterion* (paragrafo 2.2)
- La direzione della causalità è corretta

2.1.2 Archi

definizione di path definizione di open backdoor elenco di tutte le path possibili con esempio

2.2 Back door criterion

2.3 Collider Bias

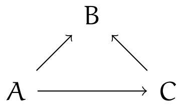

Figure 2.5: DAG:collider bias

Chapter 3

Potential outcome model

The potential outcome model defines the causal effect of an event A as the difference between the two possible states of the world, namely the world where A occurs and the one where A does not occur. For example, we would like to understand if a medicine can truly improve headaches. Let's formalize the problem by setting X as the set of patient covariates (Age , Gender,...), D as the treatment regimen (set equal to 1 if the patient takes the medicine and 0 if he takes the placebo), and Y as the number of minutes the headache persists. Under this model the causal effect of the medicine would be calculated by subtracting $Y_i|D = 1$, which we'll call Y_i^1, and $Y_i|D = 0$ which we'll call Y_i^0. These are the so called "potential outcome" because they are both *a priori* observable, but only one can be observed *a posteriori*. Let's introduce a numerical example of a random trial, where we somehow know both potential outcomes :

Age	Sex	Y_i^0	Y_i^1	δ_i
20	M	20	21	-1
20	F	15	3	12
20	M	8	10	-2
20	F	16	15	1
30	M	12	13	-1
30	F	8	5	3
30	M	2	11	-9
30	F	15	26	11

Table 3.1: Random trial knowing both potential outcomes

We can observe that δ isn't always positive, so the medicine didn't universally improve the situation, but it is seemingly clear that it had positive effects, this isn't nearly precise enough so we need to introduce some mathematical definitions :

Parameter definition

- CATE or Conditional Average Treatment Effect is defined as $E[Y_i^1 - Y_i^0|X] = E[\delta_i|X]$, so $CATE_{(M,20)} = E[\delta_i|X = (M,20)] \approx \frac{18-2}{2} = 8$, we can say that on the medicine caused a reduction on average of 8 minutes between 20 year old males.

- ATE or Average Treatment Effect is defined as $E[Y_i^1 - Y_i^0] = E[\delta_i]$, quindi $ATE = E[\delta_i] \approx \frac{18+12-2+18+3-9+11}{8}$.

- ATT o Average Treatment on the Treated is defined as $E[Y_i^1 - Y_i^0|D = 1] = E[\delta_i|D = 1]$

- ATU o Average Treatment on the Untreated is defined as $E[Y_i^1 - Y_i^0|D = 1] = E[\delta_i|D = 1]$

In this example we somehow know both potential outcomes so we can't calculate the last two quantities. It useful to make a distinction between *factual* and *counter factual* values, the former refers to the state of the world we are currently in and the latter refers to *what would have been* if the treatment were different. We can better understand the distinction between the two through the switching equation:

$$Y_i^{obs} = D_i \cdot Y_i^1 + (1 - D_i) \cdot Y_i^0 \tag{3.1}$$

We can see that Y_i^1 is either a *factual* value when $D_i = 1$ or *counter factual* when $D_i = 0$. It's obvious that in real setting, we wont ever be able to fill all the data like in table 3.1, at least half of the value will be missing and Bayesian methods will help us fill it. An example of such table could be this:

Age	Sex	Y_i^0	Y_i^1
20	M	20	?
20	F	15	?
20	M	?	10
20	F	?	15
30	M	12	?
30	F	8	?
30	M	?	11
30	F	?	26

Table 3.2: Tabella esperimento

The potential outcome model manages to transform the intractable problem of causation in to a much more manageable problem of *missing data*.

3.1 Assumptions and exclusion restriction

So far we hid a few of the assumption to avoid making this introduction unnecessary cumbersome however it is now necessary to point them out to avoid any unnecessary confusion. These are the so called *exclusion restriction* : assumption made with substantive knowledge of the subject matter, in fact any of these assumptions can be loosened if the specific case requires it, but they are considered to be most common assumption when tackling a problem of causal inference (Imbens & Rubin, 2015)

3.1.1 SUTVA

This *exclusion restriction* first proposed in (Rubin, 1980), states that :

Assumption 3.1. *The potential outcome for any unit do not vary with the treatments assigned to other units and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

(Imbens & Rubin, 2015)

This assumption can be split in two parts, the No interference component and the no hidden variation of treatment components. The former implies that any the potential outcomes for a given observation are independent from treatment of other units, it can be expressed by :

$$D_j \perp\!\!\!\perp (Y_i^0, Y_i^1) \forall i \neq j$$

For example in the headache situation we excluded *a priori* that each unit treatment wouldn't effect an other, of course this kind of assumption ought to be carefully considered by an expert on a case by case basis. This kind of assumption becomes shaky when we consider time series or pandemics where a unit treatments is very likely to impact the other chances of survival. The problem can be solved in two ways either by having a more suiting *exclusion restriction* or by defining a unit in a broader term, in the vaccination example the unit could be the nation (given no spillover between nations), instead of the individual. The second component instead implies that the treatment is homogeneous in the population, so back the headache example we cant have a medicine that has of different potency or is administered in different ways.

3.1

3.1.2 Strong ignorability

This assumption was first formalized in (Rosenbaum & Rubin, 1983) , and aimw to restrict the possible assignment mechanism (e.g.)and has three main components :

Assumption 3.2 (Individualistic Assigment)**.** *An assignment mechanism is individualistic if* $\Pr(W_i|X_i, Y_i^0, Y_i^1) = q(X_i, Y_i^0, Y_i^1)$ *for some function* $q(.)$

Assumption 3.3 (Probabilist Assigment)**.** *An assignment mechanism is probabilistic if* $\Pr(W_i|X_i, Y_i^0, Y_i^1)$ *is bounded strictly between 0 and 1* $\forall X, Y^1, Y^0, i$

This requirement is important because we need to see the counter factual in the limit, the estimation if this condition is not met is practically impossible

Assumption 3.4 (Uncofounded assigmnet)**.** *An assignment mechanism* $\Pr(W_i|X_i, Y_i^0, Y_i^1)$ *is unfounded if it does not depend on the potential outcomes* $\Pr(W_i|X_i, Y_i^0, Y_i^1) = \Pr(W_i|X_i)$ *or* $D_i \perp\!\!\!\perp (Y^0, Y^1)|X$

This assumption is equivalent to the back-door criterion this assumption is not testable (pag 261 IR)

"fan li : a tutorial on causal inference"

3.1.3 Superpopulation and finite samples

3.2 Studi randomizzati: Ruolo della randomizzazione

We often hear about double-blind randomized trials, where some of the patients are given the medicine and the other are given a placebo with neither the doctors nor the patients aware of who received what. Such trial can be expressed through a DAG in this way:

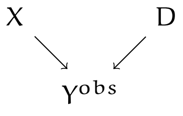

Figure 3.1: Dag per studio randomizzato

Why is this kind of trial become the golden standard for causal inference? What can such a trial truly tell us? Under the potential outcome model this kind of trial imply this independence:

$$D \perp\!\!\!\perp (Y^0, Y^1) \tag{3.2}$$

This means that the potential outcomes didn't play a role in determining the treatment regime. This **doesn't** imply that $D \perp\!\!\!\perp Y^{obs}$, in fact is patently false each time a medicine has an effect on patient. By virtue of equation 3.2, we can state that $E[Y_i^1] = E[Y_i^1|D_i = 1]$ and the same for $E[Y_i^0] = E[Y_i^0|D_i = 0]$. We can than say that:

$$\begin{aligned} ATE &= E[Y_i^1 - Y_i^0] && (3.3)\\ &= E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 0] && (3.4)\\ &= E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1] && (3.5)\\ &= E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0] && (3.6) \end{aligned}$$

In the equation 3.4 both quantities are *factual*, through the law of large number we can estimate both as means. We will define as SDO the simple difference in means in the observation:

$$SDO := \frac{1}{N_1} \sum_{i:d_i=1} y_i - \frac{1}{N_2} \sum_{i:d_i=0} y_i \overset{(N_1,N_2)\to\infty}{=} ATE$$

We can also conclude from equation 3.5 and 3.6 that in a randomized trial we have $ATE = ATT = ATU$.

3.2.1 Parte empirica

3.3 Observational studies

Observational studies are much different, researchers gather data that isn't intended to be randomized trials, were the researchers cannot intervene and impose the randomization, for two main reason :

1. Ethical or practicability concerns (EX: if we wanted to know whether smoking causes cancer)
2. The events might have already taken place (EX: if we wanted to study the effect of a certain policy)

The main difference is that $D \not\perp\!\!\!\perp (Y^0, Y^1)$, we can represent this too with a DAG:

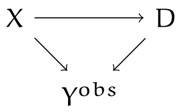

Figure 3.2: Dag per studio osservazionale

Bibliografia

Baldi, P. & Shahbaba, B. (2020). Bayesian causality. *The American Statistician* **74**, 249–257.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association* **75**, 591–593.