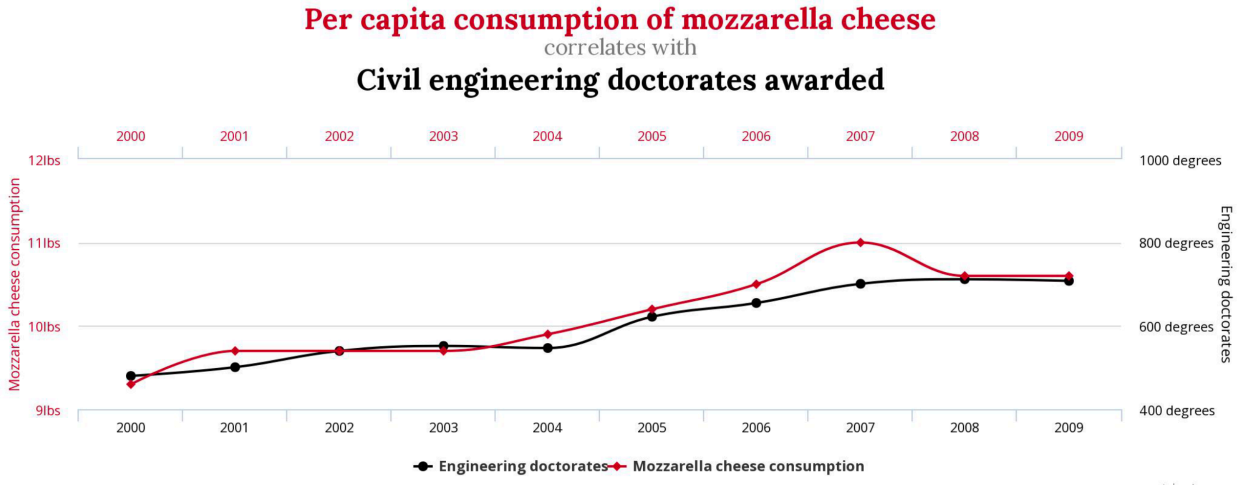
Indice

Introduzione parte II

Questa parte di corso si concentra sul condurre inferenza su i dati disponibili. In particolare, la statistica classica ci può aiutare a trovare correlazione tra i dati, ci avverte che correlazione non è causalità[1], ma non ci dice quale sia la causa di un fenomeno e come si identifica. Le risposte a queste domande non risiedono nei meri dati.

[1] Per esempio, se i lavoratori L di un'azienda sono proporzionali al fatturato F tramite una costante di proporzionalità k , allora $L = kF$, $L - kF = 0$ e $F = L/k$ sono tutte corrette, non abbiamo informazioni su quale sia il driver, se il fatturato o il numero di lavoratori.



Correlazione significa co-variabilità: variabili correlate tendono a variare assieme. La causalità implica una relazione tra due variabili di cui una è causa dell'altra. Si noti che la correlazione è condizione necessaria ma non sufficiente per la causalità.

Ceteris Paribus Condition

Prendiamo come esempio l'analisi tra il debito sottoscritto dagli studenti e i loro risultati accademici, e supponiamo di trovare una correlazione negativa. Supporre allora che il debito abbia un'influenza negativa sulla performance accademica degli studenti può essere fuorviante: il fatto che lo studente venga da una famiglia meno abbiente può infatti essere ragionevolmente la causa di ambi i fenomeni. Questo vuol dire che il confronto non è *a parità di condizioni* i.e. *ceteris paribus*. Il problema principale dell'inferenza nasce dal fatto che non possiamo confrontare i risultati accademici degli stessi studenti se si indebitassero o meno. L'econometria cerca quindi di costruire situazioni simili ad hoc, ovvero usa i dati per trovarsi il più possibile nella situazione di "a pari condizioni", in mo-

do da evitare i problemi che possono generare stime scorrette dell'effetto causale. Questo lavoro sui dati viene descritto dal *do operator*.

Do Operator

Immaginiamo di predirre Y usando X su certi dati osservazionali. In generale

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \tag{1}$$

Questa non è però una relazione causale, ma solo condizionata. Occorre dunque introdurre il do operator, condizionando quindi anziché su X su do(X). In altre parole, dovremmo calcolare $P(Y|\text{do}(X))$, che è la distribuzione condizionata di Y se fossimo in grado di intervenire nel processo di generazione dei dati e imporre $X = x$.

DAGs

Nel 2000 J. Pearl ha introdotto un chiaro e innovativo approccio grafico al tema della causalità: *Directed Acyclic Graphs*. L'obiettivo di un DAG è quello di disegnare un sistema causale per rappresentare esplicitamente tutte le cause dell'outcome d'interesse. Il modello di causalità è semplificato in quanto:

- Assume un effetto omogeneo su tutte le osservazioni
- Utilizza solo outcome osservabili e non potenziali
- Si concentra sull'effetto medio incondizionato del trattamento
- Non specifica il tipo di relazione tra le variabili

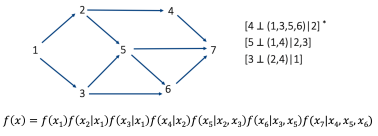
Un po' di terminologia riguardante i grafici aciclici orientati:

- Orientato: tutte le relazioni puntano da una causa a un effetto
- Aciclico: partendo da un qualunque vertice non possiamo tornare allo stesso
- Percorso: una qualsiasi sequenza di vertici orientati in una qualsiasi direzione
- Percorso diretto (o causale): un percorso in cui tutti i nodi puntano in una sola direzione

Vengono inoltre detti:

- Genitore: una causa diretta di una variabile
- Figlio: l'effetto diretto di una certa variabile
- Antenato: una causa diretta o indiretta di una certa variabile
- Discendente: un effetto diretto o indiretto di una certa variabile

DAG - Modello di Markov



Un DAG è un modello in cui le variabili dipendono solo dai propri genitori e sono indipendenti dalle altre variabili. La funzione di probabilità di massa è

$$f(x) = \Pi_{i=1}^{N} f(x_i|x_{iPA}) \tag{2}$$

Se la relazione tra A e B è mediata da C, C è detto mediator

Se A e B concausano C, C è un collider
Se A e B sono causati da C, C è un cofounder
Attenzione perché il cofounder è esattamente ciò che crea problemi nell'analisi: infatti è facile confondere la correlazione tra A e B per causalità, mentre la causa della correlazione è C. Un percorso causale può essere immediato $A \to B$ o mediato, come per esempio $A \to B \to C$. L'effetto totale di A su B è la combinazione di tutti i percorsi immediati e mediati da A a B.

L'utilizzo dei DAG permette di offrire una rappresentazione delle relazioni causali, chiarire le domande di ricerca ed evidenziare i concetti rilevanti, rendere esplicite le assunzioni dei nostri modelli, indentificare appropriatamente le variabili da inserire nell'analisi e ottenere risultati più affidabili, riducendo possibili bias. Per costruire un DAG:

- Articolare la domanda di ricerca identificando la causa e l'effetto a cui si è interessati («qual è l'effetto di A su B?»)
- Identificare altre variabili rilevanti per la relazione, come collider e mediator
- Identificare variabili cofounder
- Identificare eventuali variabili non misurate o non misurabili
- Identificare possibili processi di selezione nello studio

Negli anni sono stati creati strumenti a supporto della costruzione di un DAG come `http://www.dagitty.net/`

Paradosso di Simpson

Situazione statistica nella quale un trend o una relazione che è osservata tra diversi sottogruppi sparisce quando i gruppi sono combinati. In altre parole, dividendo i dati in gruppi, le conclusioni sono diverse rispetto a quelle derivanti da un'analisi aggregata.