Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

Analisi di mediazione: teoria e applicazioni

Relatrice: Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Laureanda: Sara Bisson
Matricola n. 2007336

Anno Accademico 2022/2023

Indice

Capitolo 1

Introduzione

Baron e Kenny (1986) definiscono il mediatore come "il meccanismo generatore attraverso il quale la variabile indipendente focale riesce a influenzare la variabile dipendente di interesse". L'aggiunta di una terza variabile all'interpretazione della relazione fra una variabile indipendente e una variabile dipendente incrementa il numero, ma anche la complessità, delle possibili relazioni fra le tre variabili. L'analisi di mediazione ha il potenziale di apportare miglioramenti sostanziali nell'ambito della sanità pubblica, contribuendo all'identificazione di strategie per prevenire comportamenti problematici e promuovere quelli vantaggiosi. La sua rilevanza si estende ampiamente anche ad altre discipline, in particolare nella sfera della psicologia, delle scienze sociali e mediche.

All'inizio degli anni '70, studiosi in vari ambiti hanno iniziato a riconoscere l'utilità del modello di mediazione per la ricerca nell'ambito dei trattamenti sanitari e della prevenzione. Nel corso del tempo, questo modello è stato impiegato per individuare variabili mediatrici, definite *endpoint* surrogati o intermediari, che fungono da indicatori precoci di patologie future. Tali variabili rivestono un ruolo importante, in quanto gli effetti dei trattamenti su di esse sono più facilmente esaminabili rispetto agli effetti sulle malattie stesse, le quali potrebbero presentare frequenze basse e richiedere un lungo periodo di tempo per manifestarsi.

Judd e Kenny (1981) hanno delineato l'impiego del modello di mediazione nell'ambito dei programmi sperimentali di promozione della salute e prevenzione delle malattie. In seguito, Baron e Kenny (1986) hanno fornito un'analisi approfondita delle variabili mediatrici nell'ambito delle scienze sociali, includendo metodi per l'analisi dei loro ef-

fetti. MacKinnon e Dwyer (1993) hanno poi descritto l'applicazione alla ricerca sulla prevenzione e hanno valutato vari aspetti statistici sottostanti all'analisi di mediazione.

L'individuazione dei legami tra cause ed effetti ha da sempre rivestito un ruolo centrale nell'orientamento con cui gli esseri umani percepiscono il proprio ambiente, poichè, come riassunto da Shultz (1982), "il concetto di causalità è indispensabile alla conoscenza umana tanto quanto i concetti di oggetto, spazio, tempo, quantità e logica".

Nonostante il concetto di mediazione possa sembrare elementare, l'analisi statistica di tali variabili rivela un notevole grado di complessità. L'obiettivo di questa tesi è quello di fornire una panoramica degli aspetti statistici, metodologici e concettuali correlati all'analisi di mediazione.

L'elaborato è strutturato come segue. Nel capitolo 2 viene delineato il modello di mediazione categoriale con la variabile dipendente dicotomica, con un approfondimento al caso in cui si renda necessaria l'analisi di più mediatori. L'argomento si conclude con un esempio tratto dalla letteratura, che contribuisce a migliorare la comprensione dell'analisi di mediazione categoriale. Nel capitolo 3 si passa al modello di mediazione continua, analizzando in particolare il modello singolo, il modello longitudinale e infine il modello multilivello. Una volta affrontata dal punto di vista teorico, l'analisi di mediazione continua viene esplorata nel capitolo 4 tramite un'applicazione a dati reali. Nel capitolo 5 vengono quindi presentate delle conclusioni che mettono in luce i risultati ottenuti durante lo sviluppo del lavoro.

Capitolo 2

Analisi di mediazione categorica

Il presente capitolo si pone l'obiettivo di approfondire il ruolo che assume il mediatore quando si è in presenza di variabili categoriali. Sebbene la mediazione sia stata ampiamente studiata nei casi di variabili continue, le applicazioni a variabili categoriali sono state meno esplorate e richiedono l'utilizzo di tecniche specifiche (MacKinnon *et al.*, 2007).

In particolare, si esamina l'analisi di mediazione su dati in cui le variabili X (variabile indipendente), M (variabile mediatrice) o Y (variabile dipendente) sono categoriche. Una possibile strategia proposta per affrontare tale scenario consiste nell'utilizzare modelli lineari quando la variabile X assume una configurazione binaria, mentre M e Y assumono valori continui. In questo caso, le tecniche standard di regressione si rivelano idonee poiché la variabile X funge esclusivamente da predittore nelle equazioni di regressione. Tuttavia, nel caso in cui il mediatore M e/o la variabile dipendente Y sono binarie, l'impiego degli approcci tradizionali risulta inadeguato poichè entrambe le variabili assumono il ruolo di variabili dipendenti.

Sono state proposte diverse soluzioni per affrontare il problema delle variabili categoriche o di una combinazione di variabili categoriche e continue nell'analisi di mediazione. Ad esempio, Winship e Mare (1983) e Muthén (1984) hanno ipotizzato l'esistenza di una variabile latente continua sottostante a qualsiasi variabile discreta o categorica, con dati categorici osservati che si manifestano attraverso una funzione di soglia. Questi approcci utilizzano distribuzioni normali o binomiali per supportare l'approccio probabilistico alle categorie osservate e si avvalgono dei minimi quadrati generalizzati per la stima

dei parametri. Sebbene teoricamente forti, tali approcci richiedono ipotesi complesse e l'implementazione pratica può risultare difficile. MacKinnon e Dwye (1993), d'altra parte, hanno utilizzato fattori di correzione della varianza per equalizzare le scale e hanno cercato di trasformare le stime dei parametri per adattarle alle diverse tipologie di variabili.

Sono stati sviluppati anche metodi di analisi di mediazione più generali e flessibili nell'ambito dell'inferenza causale. VanderWeele (2016) fornisce una panoramica autorevole di questa letteratura, in cui l'analisi di mediazione causale si basa sulla definizione chiara degli effetti diretti e indiretti di X su Y. Tali effetti possono essere stimati dai dati osservabili, a condizione che siano soddisfatte adeguate ipotesi. Esiste una vasta letteratura sulle ipotesi necessarie, sulle progettazioni di ricerca che favoriscono la loro soddisfazione e sull'analisi della sensibilità delle conclusioni alle violazioni di tali ipotesi.

In generale, esistono diverse combinazioni di variabili X, M e Y come continue o categoriche, e sono disponibili diversi modelli per adattarsi alle diverse situazioni.

2.1 Variabile dipendente categoriale

Le variabili risposta categoriali si riscontrano frequentemente in diverse tipologie di studi, quali quelli che indagano la soddisfazione del cliente, le preferenze dei consumatori, l'adesione a un trattamento, le indagini sull'incidenza di una particolare malattia e una vasta gamma di altre tematiche.
Per questo tipo di analisi è necessario utilizzare un modello che consideri alcune questioni di fondamentale importanza. In particolare, il modello deve tenere conto del fatto che i valori predetti devono rientrare nel *range* di possibili valori (Stoltzfus, 2011) - compresi tra 0 e 1 nel caso di una variabile dicotomica - e che i residui del modello non seguono più una distribuzione normale, come accade invece nei modelli lineari. Queste considerazioni sono essenziali per ottenere risultati accurati e interpretabili.

Quando si ha a che fare con una variabile dipendente dicotomica la regressione logistica è uno dei metodi statistici più utilizzati (Maalouf, 2011; Stoltzfus, 2011). Lo scopo è determinare la probabilità di un evento specifico, tenendo conto di determinati fattori. In contesti come quello medico, la regressione logistica consente di rispondere a domande cliniche fondamentali, ad esempio riguardo alla presenza di complicazioni, malattie o eventi avversi, in relazione alla presenza o assenza di specifici fattori di rischio

clinicamente rilevanti, come obesità, età, fumo, diabete, e così via.
La regressione logistica, come anche la regressione lineare, richiede il rispetto di determinate assunzioni, essenziali per assicurare che i risultati della regressione non siano affetti da distorsione. Queste sono:

1. Linearità tra la funzione logit della media della risposta e ciascuna delle variabili indipendenti continue: è importante verificare che la relazione fra queste variabili sia approssimativamente lineare nella scala logit. Nel caso in cui venisse rilevata una relazione non lineare devono essere prese in considerazione trasformazioni o modelli non lineari (De la Cruz *et al.*, 2011).

2. Mancanza di multicollinearità: la multicollinearità si verifica quando le variabili indipendenti sono altamente correlate tra loro, il che può causare instabilità nella stima dei coefficienti di regressione e rendere difficile l'interpretazione dei risultati. È necessario valutarne la presenza e, se necessario, adottare misure correttive come la rimozione o la combinazione delle variabili correlate (Bayman e Dexter, 2021).

3. Indipendenza delle osservazioni: le osservazioni devono essere raccolte in modo indipendente l'una dall'altra. In presenza di dati correlati o raggruppati, possono essere richiesti modelli statistici più avanzati, come la regressione logistica multilivello (Austin e Merlo, 2017).

4. Campione sufficientemente ampio: è importante disporre di un campione di dimensioni adeguate per evitare l'adattamento eccessivo del modello ai dati, che potrebbe portare a una scarsa generalizzazione ai nuovi dati (Perneger *et al.*, 2023).

5. Assenza di *outliers* che influenzino in modo rilevante i dati: i valori anomali possono alterare i risultati della regressione logistica. Pertanto, è importante identificarli e gestirli in modo appropriato.

Date queste assunzioni, sia $X \in R^{n \times p}$ una matrice di dati, dove n è il numero di unità statistiche e p è il numero di variabili indipendenti, e sia $y = (y_1, y_2, ..., y_n)$ un vettore di risultati binari: per ogni $x_i \in R^p$, con $i = 1 \ ... \ n$, la risposta può essere $y_i = 1$ o $y_i = 0$. Può essere quindi considerata come una variabile casuale di Bernoulli con valore atteso $E(Y_i) = \pi_i$ e distribuzione di probabilità

$$P(Y_i = y) = \begin{cases} \pi_i, & \text{se } y = 1 \\ 1 - \pi_i, & \text{se } y = 0. \end{cases} \tag{2.0}$$

La trasformazione logistica (logit) è il logaritmo dell'odds ratio ed è definita come

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i\beta, \tag{2.1}$$

dove $\beta_1, ..., \beta_p$ sono i parametri di regressione da stimare.

Quando la variabile dipendente è di natura dicotomica, è possibile calcolare l'effetto del mediatore utilizzando due approcci distinti. Tuttavia, poiché esiste un'equivalenza tra i due metodi nei modelli lineari e continui, risulta conveniente formulare il problema all'interno di un modello di questo tipo:

$$Y^* = \beta_1 + cX + e_1, \tag{2.2}$$

$$M = \beta_2 + aX + e_2, \tag{2.3}$$

$$Y^* = \beta_3 + c'X + bM + e_3. \tag{2.4}$$

La variabile dipendente $Y^* = \log\left(\frac{\pi}{1-\pi}\right)$ rappresenta la variabile latente continua sottostante che non viene osservata direttamente, ma che viene dicotomizzata in una delle categorie della variabile risposta osservata Y (che assume valori 0 o 1 per una variabile binaria). In particolare, $Y = 1$ se e solo se $Y^* > 0$.

Nel contesto del modello considerato, gli errori casuali e_i, con i=1,2,3, sono indipendenti e seguono una distribuzione logistica standard, con media zero e varianza $\frac{\pi^2}{3}$ (MacKinnon e Dwyer, 1993; MacKinnon, 2008).

I parametri utilizzati rappresentano:

- *c*: l'effetto di X su Y^*, ovvero l'`effetto totale`
- *c'*: l'effetto di X su Y^* dopo l'aggiustamento per M, ovvero l'`effetto diretto`
- *a*b*: l'effetto di M su Y^*, ovvero l'`effetto indiretto`.

Si osserva che sia il parametro c che il parametro c' rappresentano le relazioni tra la variabile indipendente X e la variabile dipendente Y. Tuttavia, c' rappresenta un effetto parziale che è stato aggiustato dal mediatore M.

L'effetto del mediatore può anche essere espresso come la differenza tra c e c', quantità che riflette quanto il cambiamento di un'unità di X influenza Y indirettamente attraverso M. L'effetto diretto, invece, rappresenta quanto della relazione tra X e Y è spiegato dal mediatore.

I parametri del modello possono essere stimati utilizzando la regressione ai minimi quadrati ordinari.

È possibile costruire intervalli di confidenza per fornire un intervallo di valori plausibili, utilizzando sia gli *standard error* basati su $\hat{a}\hat{b}$ sia quelli basati su $\hat{c} - \hat{c}$':

$$\text{LCL } (\textit{Lower confidence limit}) = \text{effetto mediatore} - z_\alpha * s_{\hat{a}\hat{b}} \tag{2.5}$$

$$\text{UCL } (\textit{Upper confidence limit}) = \text{effetto mediatore} + z_\alpha * s_{\hat{a}\hat{b}} \tag{2.6}$$

Per calcolare l'errore *standard* di $\hat{a}\hat{b}$, sono disponibili due formule. La prima formula (2.7) è la più comunemente utilizzata, mentre la seconda formula (2.8) è computazionalmente più semplice da utilizzare, in particolare quando i coefficienti e gli errori *standard* sono piccoli:

$$s_{\hat{a}\hat{b}} = \sqrt{\hat{a}^2 s_{\hat{b}}^2 + \hat{b}^2 s_{\hat{a}}^2} \tag{2.7}$$

$$s_{\hat{a}\hat{b}} = \frac{\hat{a}\hat{b}\sqrt{t_{\hat{a}}^2 + t_{\hat{b}}^2}}{t_{\hat{a}} t_{\hat{b}}}. \tag{2.8}$$

Lo *standard error* di $\hat{c} - \hat{c}$' è invece più complicato da calcolare, in quanto richiede diversi passaggi (Freedman, 1992).

Nella regressione logistica è importante considerare che la scala per la variabile dipendente Y^* non è direttamente osservabile. Per fissare la scala di Y^*, viene impostata una varianza residua specifica pari a $\frac{\pi^2}{3}$ (MacKinnon e Dwyer, 1993; MacKinnon, 2008). Questa scelta garantisce che le varianze residue siano fissate, ma comporta una differenza nella scala di Y^* tra i diversi modelli. Pertanto, è fondamentale tenere presente che i metodi che utilizzano la differenza tra le stime c e c' non sono equivalenti a quelli che

utilizzano il prodotto delle stime a e b per stimare l'effetto di mediazione.

Una soluzione a questo problema è rendere la scala equivalente tra le equazioni mediante la standardizzazione dei coefficienti di regressione prima del calcolo della stima della mediazione (Winship e Mare, 1983).

Ad esempio, la varianza di Y^* nelle equazioni (2.2) e (2.4) è descritta rispettivamente da:

$$\hat{\sigma}^2_{Y^*} = \hat{c}^2\hat{\sigma}^2_X + \frac{\pi^2}{3}, \tag{2.9}$$

$$\hat{\sigma}^2_{Y^*} = \hat{c'}^2\hat{\sigma}^2_X + \hat{b}^2\hat{\sigma}^2_M + 2\hat{c'}\hat{b}\hat{\sigma}_{XM} + \frac{\pi^2}{3}. \tag{2.10}$$

Si possono quindi dividere $\hat{c}$ per il valore di $\hat{\sigma}_{Y^*}$ ottenuto dall'equazione (2.9), e $\hat{c}$' e $\hat{b}$ per $\hat{\sigma}_{Y^*}$, ottenuto dall'equazione (2.10). Dopo aver diviso gli *standard error* per la radice quadrata della varianza, l'effetto di mediazione $\hat{a}\hat{b}$ si ottiene rapportandolo al suo *standard error* e confrontandolo con una distribuzione normale *standard* per testarne la significatività.

Un metodo alternativo consiste nell'utilizzare $\hat{c}_{corretto}$ (2.11) con $\hat{\sigma}^2_{resM}$ varianza residua dell'equazione (2.3) per trasformare il coefficiente $\hat{c}$ nella stessa scala del coefficiente $\hat{c}$':

$$\hat{c}_{corretto} = \hat{c}\sqrt{1 + \frac{\hat{b}^2\hat{\sigma}^2_{resM}}{\frac{\pi^2}{3}}}\ . \tag{2.11}$$

Se il mediatore è una misura continua, la stima del parametro a dalla regressione ordinaria non richiede la standardizzazione: il prodotto tra a e b standardizzato corrisponde all'effetto di mediazione.

L'effetto mediato e il suo errore *standard* forniscono una misura di quanto la mediazione contribuisca all'effetto totale. Tuttavia, questa misura da sola non fornisce informazioni sulla dimensione relativa della mediazione rispetto all'effetto diretto.

Per comprendere meglio la magnitudine della mediazione, si possono utilizzare due misure aggiuntive. La prima è la percentuale dell'effetto totale che viene mediato, ovvero $\frac{\hat{a}\hat{b}}{\hat{a}\hat{b}+\hat{c'}}$.

La seconda misura, utilizzabile solo nel caso in cui l'effetto diretto sia diverso da zero, è invece il rapporto tra l'effetto mediato e l'effetto diretto. L'effetto mediato è approssimativamente $\frac{\hat{a}\hat{b}}{\hat{c'}}$ dell'effetto diretto.

2.2 Modello di mediazione multipla

Spesso viene ipotizzato più di un processo mediatore nella relazione fra una variabile indipendente e una variabile dipendente. Quando sono ipotizzati più mediatori, generalmente conviene utilizzare un modello di mediazione multipla piuttosto che diversi modelli singoli. Il principale vantaggio di questo modello è che consente di stimare gli effetti indiretti specifici condizionatamente alla presenza di altri mediatori. Ciò permette di ottenere stime dei parametri più accurate, poiché tengono conto delle informazioni provenienti dagli altri mediatori. Il modello di mediazione multipla è la base teoretica di molti programmi di prevenzione in ambito medico: l'analisi dettagliata dei contributi dei mediatori alle variazioni della variabile dipendente potrebbe mostrare i mediatori critici e aiutare a risolvere discrepanze fra gli studi.

Essendo un'estensione diretta del modello di mediazione singola (2.1), le equazioni (2.3) e (2.4) possono essere modificate al fine di includere più di un mediatore. Ad esempio, quando la variabile X ha effetto su due mediatori, che hanno entrambi effetto sulla variabile Y, vengono utilizzate quattro equazioni di regressione:

$$Y^* = \beta_1 + cX + e_1, \tag{2.12}$$

$$M_1 = \beta_2 + a_1X + e_2, \tag{2.13}$$

$$M_2 = \beta_3 + a_2X + e_3, \tag{2.14}$$

$$Y^* = \beta_4 + c'X + b_1M_1 + b_2M_2 + e_4. \tag{2.15}$$

In questo modello M_1 è il primo mediatore, M_2 è il secondo mediatore, c' è il parametro che lega X a Y^* aggiustando per i due mediatori, b_1 è il parametro che lega il primo mediatore a Y^* aggiustando per X e per M_2 e b_2 è il parametro che lega il secondo mediatore a Y^* aggiustando per X e per M_1. Si nota che, come per il modello di mediazione singola, sia c che c' legano la variabile indipendente a quella dipendente, ma c' è aggiustato dall'effetto dei mediatori.

Le interpretazioni dell'effetto totale e dell'effetto diretto restano immutate rispetto a quelle del modello di mediazione singola.

Gli effetti dei due mediatori sono invece rappresentati rispettivamente da $\hat{a}_1\hat{b}_1$ e da $\hat{a}_2\hat{b}_2$. Ogni effetto di mediazione del tipo $\hat{a}_i\hat{b}_i$ (i=1,2) prende il nome di effetto indiretto specifico, per distinguerlo dall'effetto totale di mediazione, che è la somma di questi effetti specifici. Quando il numero di mediatori aumenta, aumenta anche il numero di effetti specifici di mediazione.

Gli intervalli di confidenza e gli *standard error* si individuano come nel modello con un singolo mediatore, con un adeguato cambio di notazione. La soluzione delta multivariata per l'errore *standard* dell'effetto di mediazione totale $\hat{a}_1\hat{b}_1 + \hat{a}_2\hat{b}_2$ equivale a

$$s_{\hat{a}_1\hat{b}_1+\hat{a}_2\hat{b}_2} = \sqrt{\hat{b}_1^2 s_{\hat{a}_1}^2 + \hat{a}_1^2 s_{\hat{b}_1}^2 + \hat{b}_2^2 s_{\hat{a}_2}^2 + \hat{a}_2^2 s_{\hat{b}_2}^2 + 2\hat{a}_1\hat{a}_2 s_{\hat{b}_1\hat{b}_2}}, \tag{2.16}$$

dove $s_{\hat{b}_1\hat{b}_2}$ è la covarianza fra $\hat{b}_1$ e $\hat{b}_2$.

Come descritto nel paragrafo precedente, l'effetto del singolo mediatore può essere testato per la significatività statistica dividendo la stima del suo effetto per il suo *standard error* e confrontandolo con la distribuzione normale *standard*.
L'uguaglianza degli effetti di mediazione può invece essere testata utilizzando la quantità $s_{\hat{a}_1\hat{b}_1-\hat{a}_2\hat{b}_2}$, formula che si ottiene come la (3.4) ma sottraendo l'elemento $2\hat{a}_1\hat{a}_2 s_{\hat{b}_1\hat{b}_2}$ al posto di sommarlo.

Il metodo più usato per valutare la mediazione è stato descritto da Baron e Kenny (1986). Il metodo degli *step* causali consiste in una serie di *test* statistici per verificare la relazione fra variabili. Le limitazioni dei seguenti passaggi mostrano diversi aspetti unici del modello di mediazione multipla.

1. La variabile indipendente X deve avere un effetto sulla variabile dipendente Y^* (coefficiente $\hat{c}$ in (2.12)).
   L'obiettivo di questo primo *test* è quello di stabilire se c'è un effetto da mediare. Se l'effetto non è statisticamente significativo, l'analisi si ferma. È in realtà possibile che la relazione fra la X e la Y sia non statisticamente significativa, tuttavia può ancora esserci rilevante mediazione. Questo succede in casi chiamati di mediazione inconsistente e questi tipi di modelli possono essere complicati per la mediazione

multipla, poiché alcuni effetti di mediazione potrebbero avere segni diversi fra di loro e dall'effetto diretto.

2. La variabile indipendente X deve avere un effetto sia sul primo mediatore M_1 (coefficiente $\hat{a}_1$ in (2.13)) che sul secondo mediatore M_2 (coefficiente $\hat{a}_2$ (2.14)).
   Questo *test* richiede che la variabile indipendente sia statisticamente legata ai mediatori. Anche in questo caso è possibile che i coefficienti $\hat{a}_i$ risultino non significativi per uno o più effetti di mediazione, implicando che la variabile di mediazione corrispondente non sia un vero mediatore nella relazione fra X e Y.

3. Il mediatore deve avere un effetto sulla variabile dipendente Y^* quando la variabile indipendente X è controllata (coefficienti $\hat{b}_1$ e $\hat{b}_2$ in (2.15)).
   Questo *test* è equivalente a verificare se aggiungere lo specifico mediatore modifica la relazione fra la X e la Y. Se il mediatore non è legato alla Y, l'effetto di X sul mediatore non può essere trasmesso alla variabile dipendente. Una limitazione importante di questo *test* è che risulterà significativo ogniqualvolta uno qualsiasi dei mediatori in un modello di mediazione multipla abbia una relazione significativa con la variabile dipendente. Pertanto, non sarà possibile distinguere gli effetti mediati tramite ciascun mediatore.

4. L'effetto diretto deve essere non significativo (coefficiente $\hat{c}$' in (2.15)).
   Nella letteratura che tratta il metodo degli *step* causali vi sono opinioni differenti riguardanti quest'ultimo punto. Alcuni metodi richiedono la mediazione totale, ovvero che la variabile indipendente non abbia un effetto significativo sulla variabile dipendente quando il mediatore è controllato. In questo caso l'effetto diretto deve essere non significativo. Altri metodi consentono invece la mediazione parziale, ossia che l'effetto della variabile indipendente sulla variabile dipendente sia maggiore quando non si considera il mediatore ($\hat{c}$ maggiore di $\hat{c}$'). Questi metodi permettono che $\hat{c}$' sia significativo, e ciò risulta ragionevole considerando che la mediazione completa è irrealistica in molte aree di ricerca.

In generale, vi sono diversi limiti nell'applicazione del metodo degli *step* causali. In primo luogo, le stime degli effetti di mediazione $\hat{a}_i\hat{b}_i$ non sono disponibili direttamente tramite questo metodo, anche se possono essere ottenute combinandolo con altri approcci. L'integrazione di più mediatori con il metodo dei passi causali risulta generalmente difficile, in quanto è possibile ottenere solo l'effetto di mediazione totale.

Inoltre, studi di simulazione hanno dimostrato che la capacità di rilevare effetti mediati può essere molto bassa (MacKinnon *et al.*, 2002). L'approccio a passi causali di Baron e Kenny richiede infatti circa 21.000 soggetti per poter rilevare adeguatamente un effetto quando le dimensioni degli effetti $\hat{a}$ e $\hat{b}$ sono di debole entità e l'intera relazione tra X e Y è mediata (Fritz, 2007).
Un altro limite importante di questo metodo è la necessità di un effetto totale $\hat{c}$ significativo affinché l'analisi di mediazione possa procedere. Il requisito di una relazione complessiva significativa tra X e Y rappresenta la differenza principale tra l'approccio a passi causali e altri metodi per testare la mediazione: alcuni ricercatori l'hanno considerato come un *test* perfetto della relazione, nonostante si tratti di un *test* statistico soggetto a errori. Quando l'effetto totale non è significativo, si tratta, come menzionato in precedenza, di modelli inconsistenti. Con l'aumentare del numero di mediatori, il numero di possibili combinazioni di effetti di mediazione consistenti e inconsistenti aumenta. Di conseguenza, la richiesta di effetto totale significativo potrebbe risultare incorretta per alcuni modelli. Sono presenti diverse situazioni in cui esiste una mediazione significativa ma l'effetto complessivo di X su Y non è significativo. Ad esempio, si consideri un caso in cui il segno dell'effetto mediato differisce dal segno dell'effetto diretto, causando la relazione complessiva tra X e Y ad essere nulla.

Nonostante questi limiti, l'approccio degli *step* causali è il più comunemente utilizzato per verificare la significatività della mediazione grazie al chiaro collegamento concettuale fra le relazioni causali e i *test* statistici precedentemente menzionati.

2.3 Un esempio

L'analisi di mediazione per il caso di una variabile dipendente binaria viene illustrata nel caso di un progetto di prevenzione svolto nelle scuole (MacKinnon *et al.*, 1991), noto come Progetto di Prevenzione del Midwest (MPP). Il MPP rappresenta un'iniziativa di prevenzione svolta su più comunità, mirata a ritardare l'insorgenza dell'uso di droghe *"gateway"*, come alcol, tabacco e marijuana. Il progetto coinvolge varie realtà, tra cui programmi scolastici, genitoriali, di organizzazione della comunità, di comunicazione di massa e di politiche sanitarie.

I dati utilizzati in questo esempio sono stati raccolti presso otto scuole, di cui quattro sono state assegnate casualmente per ricevere l'intervento. I partecipanti sono stati

inizialmente sottoposti a una misurazione nell'autunno del 1984 e sono poi stati seguiti nel corso del tempo. L'intervento consisteva in dieci sessioni educative finalizzate allo sviluppo delle abilità per resistere all'uso di sostanze.

Una misura di autovalutazione relativa alla guida sotto l'influenza di droghe o alcol è stata aggiunta nell'autunno del 1988, quando gli studenti erano ormai abbastanza grandi da guidare e dopo che circa la metà di essi aveva partecipato al programma di prevenzione.
L'obiettivo è indagare come e se l'uso intenso di droghe funge da mediatore tra l'effetto del programma di prevenzione, misurato nel 1987, e la propensione alla guida dopo l'uso di sostanze, misurata l'anno successivo. Il costrutto mediante il quale si esamina l'uso intenso di droghe è stato creato attraverso la standardizzazione e la combinazione delle misure relative all'uso di marijuana nell'ultimo mese, al numero di volte in cui si è stati ubriachi nell'ultimo mese e al numero di bevande consumate quando si beve alcol.

Alla misurazione dell'autunno del 1988, sono state ottenute le seguenti stime dei parametri di regressione logit:

$$\hat{Y}^* = -0.2140X_P + 0.5695X_A + \frac{\pi^2}{3}, \tag{2.17}$$

$$\hat{M} = -0.1397X_P + 0.1234X_A, \tag{2.18}$$

$$\hat{Y}^* = -0.1812X_P + 0.2976M + 0.5253X_A + \frac{\pi^2}{3}, \tag{2.19}$$

dove X_P denota la variabile indipendente, che rappresenta l'effetto del programma di prevenzione, mentre X_A svolge il ruolo di covariata e denota l'anno scolastico che il soggetto sta frequentando.

La stima dell'effetto di mediazione, calcolata come $\hat{c} - \hat{c}$', pari a -0.033, non equivale a $\hat{a}\hat{b}$, che risluta pari a -0.042. I valori negativi di queste stime suggeriscono come l'uso intenso di droghe riduca l'effetto positivo del programma di prevenzione sulla guida dopo l'uso di sostanze. Prima della standardizzazione $\hat{a}\hat{b}$ presenta uno *standard error* $s_{\hat{a}\hat{b}} = 0.02$ e un intervallo di confidenza al 90% compreso fra -0.01 e -0.07, il che suggerisce la presenza di una mediazione significativa.

Rifacendosi alle equazioni (2.9) e (2.10) e adattandole alla presenza di una covariata, la varianza di Y^* può essere stimata come segue:

$$\hat{\sigma}^2_{Y^*} = (-0.2140)^2(0.244) + (0.5695)^2(0.245) + 2(-0.2140)(0.5695)(-0.117) + \frac{\pi^2}{3}$$
$$= 1.12,$$

$$\hat{\sigma}^2_{Y^*} = (-0.1812)^2(0.244) + (0.2976)^2(0.867) + 2(-0.1812)(0.2976)(-0.048) + (0.5253)^2(0.245)$$
$$+ 2(-0.1812)(0.5253)(-0.117) + 2(0.2976)(0.5253)(0.047) + \frac{\pi^2}{3}$$
$$= 1.19.$$

Effettuando la divisione di ciascuna stima di regressione per la deviazione standard della variabile dipendente nel relativo modello, si ottengono le seguenti stime dei parametri (i coefficienti relativi all'anno scolastico non sono presentati):

$$\hat{Y} = -0.2020X_P, \tag{2.20}$$

$$\hat{Y} = -0.1660X_P + 0.2720M. \tag{2.21}$$

Le stime degli effetti di mediazione $\hat{c} - \hat{c}' = -0.036$ e $\hat{a}\hat{b} = -0.038$ risultano simili e $s_{\hat{a}\hat{b}} = 0.02$, con intervallo di confidenza (-0.005, -0.071) per la stima $\hat{a}\hat{b}$, suggerisce ancora la presenza di mediazione significativa.
In conclusione, variazioni nell'utilizzo di droghe pesanti nell'autunno del 1987 spiegano circa il 19% ($\frac{\hat{a}\hat{b}}{\hat{a}\hat{b}-c'} = \frac{-0.038}{-0.038-0.166}$) dell'effetto totale che esercita il programma di prevenzione sulla guida sotto l'influenza di alcol o altre droghe un anno dopo.

Capitolo 3

Analisi di mediazione continua

Questo capitolo si focalizza sulla comprensione e l'approfondimento dell'analisi di mediazione statistica con esclusiva considerazione delle variabili continue. Viene affrontato in modo dettagliato il modello di mediazione singolo, sfruttando l'utilizzo dei modelli di regressione lineare e la specificazione delle relative assunzioni, nonché le tecniche per la verifica dell'esistenza di una mediazione significativa (Baron e Kenny, 1986; Kraemer *et al.*, 2008). Successivamente, si esplorano gli aspetti relativi all'analisi di modelli di tipo longitudinale, evidenziando i punti di forza e le limitazioni di tale approccio (Gollob e Reichardt 1991; Kenny, 1979), per poi concludere con il modello multilivello, particolarmente adatto quando i dati vengono raccolti a più livelli (Kreft, 1996; Barcikowski, 1981).

L'obiettivo principale di questo capitolo è fornire una comprensione approfondita dei concetti e delle tecniche fondamentali necessarie per condurre un'analisi di mediazione continua accurata.

3.1 Modello singolo

Il modello singolo rappresenta l'approccio più elementare nell'ambito dell'analisi di mediazione continua. Al fine di investigare le relazioni fra le variabili, vengono impiegate le seguenti tre formule essenziali di regressione:

$$Y = \beta_1 + cX + e_1, \tag{3.1}$$

$$M = \beta_2 + aX + e_2, \tag{3.2}$$

$$Y = \beta_3 + c'X + bM + e_3. \tag{3.3}$$

La distinzione fondamentale rispetto al modello utilizzato quando la variabile Y è categoriale (Sezione 2.1) consiste nel fatto che, in questo caso, la variabile dipendente è direttamente rappresentata dalla variabile risposta osservata Y, e non da una variabile latente. Come nel modello precedentemente illustrato, la variabile indipendente X è legata al mediatore M, che a sua volta è legato alla variabile dipendente Y. L'effetto diretto è rappresentato dal parametro c', l'effetto totale da c e l'effetto indiretto, o effetto mediato, da ab.

Un'altra differenza rispetto al modello di mediazione categoriale è che, quando la variabile dipendente è continua, l'effetto mediato ab è uguale a c - c'. Come conseguenza diretta, l'effetto totale c può essere decomposto in un effetto diretto e un effetto indiretto, $c = ab + c$'. La ragione alla base della quantità di mediazione ab risiede nella dipendenza della mediazione dal grado in cui la variabile indipendente X influisce sul mediatore (parametro a) e dal grado in cui il mediatore influisce sulla variabile dipendente Y (parametro b). Tuttavia, l'uguaglianza fra ab e c - c' non è valida quando il campione è diverso per le tre equazioni di regressione. Ciò può accadere, ad esempio, se per alcuni soggetti non viene registrata la variabile mediatrice.

Anche nel modello di mediazione continua è possibile stimare i parametri delle tre equazioni di regressione mediante l'utilizzo dei minimi quadrati ordinari. Gli *standard error* di $\hat{a}\hat{b}$ possono essere calcolati utilizzando le due formule precedentemente illustrate (2.7) e (2.8), mentre lo *standard error* di $\hat{c} - \hat{c}$' può essere ottenuto come

$$s_{\hat{c}-\hat{c}'} = \sqrt{s^2_{\hat{c}} + s^2_{\hat{c}'} - 2r s_{\hat{c}} s_{\hat{c}'}}, \tag{3.4}$$

dove la covarianza $r s_{\hat{c}} s_{\hat{c}'}$ è l'errore quadratico medio (varianza stimata di e_3 in (3.3)) diviso per il prodotto fra la numerosità campionaria e la varianza della variabile indipendente ($\frac{MSE}{n * s^2_X}$).

Frequentemente, i valori ottenuti dalle equazioni (2.7) e (3.4) sono molto simili, tuttavia, la prima equazione risulta preferibile poiché è più semplice da implementare e può

essere generalizzata anche a modelli più complessi.

Le stime degli effetti e i relativi *standard error* possono quindi essere utilizzati per calcolare gli intervalli di confidenza per l'effetto di mediazione (2.5, 2.6).

3.1.1 Verifica dell'effetto di mediazione

Spesso si è interessati a verificare se un effetto di mediazione osservato è significativamente diverso da zero. Esistono vari metodi per testare questa ipotesi, alcuni dei quali sono presentati di seguito:

- *Intervallo di confidenza al (1-α)%*: se il valore zero non rientra nell'intervallo di confidenza dell'effetto di mediazione, allora questo è considerato statisticamente significativo al livello di significatività α.

- *Confronto con la distribuzione normale*: viene calcolato il rapporto tra il valore stimato dell'effetto di mediazione e il suo errore *standard* e quindi confrontato con la distribuzione normale *standard*. Se il valore assoluto del rapporto supera 1.96, allora l'effetto di mediazione è considerato significativamente diverso da zero a livello di significatività 0.05.

- *Passi per stabilire la mediazione*: il metodo degli *step* causali di Baron e Kenny (1986), descritto nel capitolo precedente (Sezione 2.2), consiste in quattro passi volti a verificare se la variabile indipendente X ha un effetto sulla variabile dipendente Y ($\hat{c}$ significativo) e sul mediatore ($\hat{a}$ significativo), se il mediatore ha un effetto significativo sulla variabile dipendente Y quando la X è controllata ($\hat{b}$ significativo) e se non c'è effetto diretto ($\hat{c}$' non significativo).

- *Framework di mediazione*: L'approccio di MacArthur, successivamente descritto da Kraemer *et al.* (2008), è simile al metodo degli *step* causali di Baron e Kenny (1986), con la differenza principale che non cerca di specificare preventivamente i meccanismi di mediazione sottostanti, ma si basa sulla raccolta empirica dei dati per identificare le relazioni causali. Per prima cosa il modello sostiene che le relazioni non lineari tra le variabili sono considerate mediazioni solo se esiste una relazione fra X e M. In secondo luogo, l'esistenza di un'interazione fra M e Y viene inclusa nel modello e presa come evidenza di mediazione. Infine, affinchè una

variabile funga da mediatore, è necessaria l'evidenza di un cambiamento nel mediatore precedente al cambiamento della variabile dipendente. Quindi i due criteri necessari (ma non sufficienti) per stabilire una relazione causale fra due variabili sono la precedenza temporale e l'associazione, e l'obiettivo di questo approccio è di generare ipotesi su un possibile ruolo causale da testare in studi futuri.

- *Test confermativo di mediazione completa*: l'approccio confermativo descritto da James *et al.* (2006) si basa sulla distinzione fra il modello di mediazione completa e quello di mediazione parziale. Gli autori sostengono che il primo modello da testare dovrebbe essere quello di mediazione completa poichè è più parsimonioso, e questo consiste nel testare se il coefficiente $\hat{c}$' è statisticamente significativo. Il modello di mediazione completa è di fondamentale importanza per assicurarsi che l'effetto di mediazione osservato sia adeguatamente spiegato dalla variabile mediatrice e che non sia influenzato da altre variabili non considerate nel modello. Il *test* si svolge in due fasi: inizialmente si ipotizza un modello completo o parziale, testando prima quello completo. Successivamente, vengono valutati i coefficienti $\hat{a}$ e $\hat{b}_{\text{nonaggiustato}}$, ovvero la relazione fra M e Y non aggiustata. Se il coefficiente $\hat{c}$' è significativo allora c'è evidenza di mediazione parziale e non completa. Un aspetto importante di questo metodo è che il tipo di mediazione va specificato prima dell'inizio dello studio.

3.1.2 Assunzioni del modello

Ogni equazione di regressione utilizzata nell'analisi di mediazione (3.1, 3.2, 3.3) richiede una serie di assunzioni. Tra queste, le più importanti sono le seguenti:

- *Forma funzionale corretta*: le equazioni presuppongono relazioni lineari fra le variabili coinvolte, dove un cambio di un'unità nella variabile indipendente produce un certo cambio nella variabile dipendente. In questi modelli è comunque possibile gestire relazioni non lineari tramite l'applicazione di opportune trasformazioni e la specificazione di variabili indipendenti che riflettano tali relazioni. Inoltre, le relazioni fra le variabili sono additive e quindi le variabili non interagiscono fra loro. Tuttavia, nel contesto del modello di mediazione singola, un effetto rilevante è rappresentato dall'interazione tra la variabile indipendente X e il mediatore M

nell'equazione (3.3). Questo aspetto è cruciale poiché stabilisce se la relazione tra la variabile X e la Y varia in base ai diversi livelli di M.

- *Nessuna influenza omessa*: si assume che le equazioni di regressione riflettano accuratamente il modello sottostante e che quindi nessuna variabile importante sia stata omessa. Il modello assume anche che tutti i partecipanti coinvolti in uno studio di ricerca abbiano le stesse relazioni di mediazione, ovvero che non vi siano sottogruppi di soggetti con processi di mediazione differenti. Tuttavia, nella realtà dei dati empirici, è improbabile che le tre variabili del modello singolo siano le uniche variabili rilevanti, e di conseguenza, l'effetto di mediazione potrebbe variare da gruppo a gruppo (ad esempio, potrebbe differire tra maschi e femmine).

- *Misurazioni accurate*: le variabili X, Y e M devono essere misure valide e affidabili. Gli errori di misurazione possono risultare particolarmente critici nell'analisi di mediazione, poichè errori nel mediatore possono portare a effetti attenuati nella relazione fra la variabile dipendente e il mediatore (Hoyle e Kenny, 1986). Per affrontare questo problema, una soluzione consiste nel specificare modelli di misurazione per i costrutti in modo tale che un costrutto latente sia ipotizzato essere misurato da diversi indicatori fallibili.

- *Residui adeguati*: si assume che i residui nelle equazioni di regressione siano incorrelati con la variabile indipendente in ogni equazione e incorrelati fra le equazioni, indipendenti fra di loro e omoschedastici a ogni valore della variabile indipendente.

- *Precedenza temporale*: nel modello di mediazione singola si assume un ordine temporale delle variabili tale che X preceda M, che a sua volta precede Y. Questo aspetto rende complessa l'applicazione della mediazione a dati trasversali, poiché generalmente non si dispone di informazioni sulla precedenza temporale.

- *Distribuzione normale*: si assume generalmente che le variabili X, Y e M abbiano una distribuzione normale. Se X è una variabile binaria, i metodi utilizzati in questo capitolo rimangono validi, mentre se Y è binaria bisogna utilizzare il modello di mediazione categoriale descritto nel capitolo precedente (Sezione 2.1). Quando le variabili di mediazione non seguono una distribuzione normale, è opportuno utilizzare metodi di simulazione che non fanno assunzioni sulla distribuzione delle variabili. Anche il prodotto $\hat{a}\hat{b}$ viene assunto normale.

3.2 Modello longitudinale

Le misure ripetute costituiscono un elemento di notevole rilevanza nel miglioramento dell'interpretazione dei processi di mediazione, poiché consentono di analizzare non solo le variazioni tra gli individui, ma anche i cambiamenti temporali che si verificano all'interno degli stessi. Questa sezione fornisce una panoramica sulle informazioni aggiuntive disponibili per i modelli di mediazione applicati a dati longitudinali.

I dati longitudinali forniscono maggiori informazioni riguardo alla precedenza temporale di X, Y e M. Nel contesto del modello di mediazione singola, in cui tutte le variabili vengono misurate contemporaneamente, l'ordine temporale viene stabilito sulla base della teoria o di studi precedenti. I dati longitudinali invece consentono di esaminare se i cambiamenti della variabile M precedono quelli della variabile Y.

Un ulteriore vantaggio derivante dall'utilizzo di dati longitudinali è la possibilità di analizzare sia le relazioni trasversali, ovvero le relazioni tra variabili rilevate nello stesso momento, sia le variazioni che si verificano all'interno di un singolo individuo nel corso del tempo. Per quanto riguarda le relazioni trasversali, le stime degli effetti si basano sulle differenze tra gli individui. Le relazioni longitudinali, invece, consentono di analizzare le relazioni tra variabili in ogni momento di rilevazione, oltre a esaminare i cambiamenti che avvengono tra le diverse rilevazioni nel tempo. Sebbene le relazioni longitudinali siano generalmente di maggiore interesse, esistono situazioni in cui le relazioni trasversali possono risultare più importanti da investigare.

Un terzo beneficio consiste nel fatto che i dati longitudinali tengono conto di possibili spiegazioni alternative per gli effetti mediati osservati tra variabili trasversali. Una spiegazione alternativa per una relazione trasversale osservata potrebbe essere l'esistenza di una variabile omessa che ne giustifica l'associazione. Tuttavia, i cambiamenti osservati nell'individuo nel corso del tempo eliminano queste possibilità dovute alle differenze statiche tra gli individui, poiché ciascun individuo funge da proprio controllo. Ad esempio, fattori biologici quali la genetica rappresentano spiegazioni improbabili per le relazioni longitudinali, poiché tali variabili non subiscono cambiamenti nel tempo.

Gollob e Reichardt (1991) hanno identificato tre limitazioni intrinseche all'utilizzo di dati trasversali per analizzare relazioni longitudinali. In primo luogo, il tempo necessario affinchè le variabili esercitino i loro effetti rappresenta un fattore cruciale: quando le va-

riabili vengono misurate simultaneamente, potrebbe non trascorrere il tempo sufficiente affinché la variabile X influenzi la variabile M o la variabile M influenzi la variabile Y. Inoltre, le variabili possono avere effetti su se stesse nel tempo, ossia possono essere correlate in momenti successivi. Infine, la dimensione dell'effetto mediato dipende dal ritardo temporale tra le misurazioni delle variabili coinvolte: l'effetto indiretto di X su Y potrebbe variare significativamente se le misurazioni delle variabili cambiano in tempi molto ravvicinati rispetto a lunghi periodi di tempo.
Queste limitazioni, secondo Gollob e Reichardt (1991), possono portare a distorsioni nelle stime degli effetti mediati quando si utilizzano dati trasversali per analizzare relazioni longitudinali.

Per ovviare a questi limiti, Gollob e Reichardt (1991) hanno proposto l'impiego di un modello longitudinale latente per dati trasversali, in cui le misurazioni in un momento temporale precedente sono trattate come variabili latenti, cioè variabili non direttamente osservate. In questo approccio, si riduce il numero di parametri sconosciuti assumendo che le relazioni longitudinali tra le variabili siano conosciute e che le varianze siano uguali in ogni momento. Tuttavia, è importante notare che queste assunzioni possono risultare spesso irrealistiche per questo tipo di modello. Pertanto, Gollob e Reichardt (1991) hanno utilizzato tale approccio solamente a fini illustrativi, per evidenziare le difficoltà che si riscontrano esaminando le vere relazioni con dati trasversali, suggerendo invece l'adozione di modelli longitudinali alternativi, che verranno descritti in seguito.

L'utilizzo di misure ripetute introduce diversi nuovi concetti. Un primo concetto cruciale è quello della stabilità, che si riferisce alla consistenza nel tempo della media di una misura, come inizialmente descritto da Kenny (1979). Altre definizioni di stabilità richiedono la presenza di trend stabili o stabilità periodica di un processo, mentre altre ancora includono una stabilità rigorosa, in cui gli individui non mostrano alcun cambiamento nel corso del tempo, una stabilità lineare, in cui si osserva un trend lineare che varia tra le persone nel tempo, e una stabilità monotona, che riflette la capacità di un individuo di mantenere le stesse relazioni relative nel tempo (Burr e Nesselroade, 1990). L'informazione sulla stabilità ottenuta da dati longitudinali viene valutata misurando la dipendenza tra le diverse occasioni di misurazione.

Il secondo concetto fondamentale riguarda la stazionarietà, che si riferisce al grado

in cui le relazioni tra le variabili rimangono costanti nel tempo (Kenny, 1979). Tale concetto presuppone che il processo che genera i dati non subisca variazioni temporali, il che risulta poco probabile nei processi reali, specialmente in studi di lunga durata. L'informazione sulla stazionarietà può essere ottenuta dai dati longitudinali valutando l'invarianza delle relazioni tra le diverse misurazioni nel corso del tempo.

Il terzo concetto importante è quello di equilibrio, strettamente collegato alla stabilità e alla stazionarietà. Dwyer (1983) ha definito un sistema in equilibrio quando vi è una stabilità temporale nella distribuzione di covarianza e varianza tra le variabili. Nel contesto della mediazione, le relazioni tra le variabili X, M e Y devono aver raggiunto un certo equilibrio durante il periodo di raccolta dati per ottenere delle stime accurate delle loro relazioni. Nei dati trasversali, si assume che l'equilibrio sia stato raggiunto al momento delle misurazioni. Tuttavia, dai dati longitudinali si possono ottenere informazioni riguardanti l'equilibrio esaminando la somiglianza delle relazioni nel modello attraverso multiple misurazioni nel tempo.

3.2.1 Modelli di regressione a due ondate

Quando si lavora con dati raccolti in due momenti distinti, un approccio comunemente utilizzato è l'analisi della differenza tra le misurazioni. Questo metodo risulta particolarmente utile quando vi è una manipolazione sperimentale tra il pre-*test* e il post-*test*, poiché consente di esprimere il modello in termini del cambiamento delle variabili M e Y prima e dopo la manipolazione. Il processo di analisi si articola come segue: in primo luogo, si effettua il calcolo delle variazioni nel tempo per ciascuna variabile misurata (indipendente, mediatrice e dipendente), ottenendo i cambiamenti nei punteggi tra le misurazioni successive e quelle precedenti. Successivamente, si calcola il cambiamento nella variabile mediatrice e si esamina come questa variazione influenzi l'effetto della variabile indipendente sulla variabile dipendente.
Questo approccio risulta particolarmente rilevante negli studi che esplorano gli effetti di interventi o trattamenti nel corso del tempo, poiché permette di valutare l'interazione tra le variabili durante le diverse misurazioni.

Una seconda opzione è l'analisi dei punteggi di cambiamento corretti per le variabili X, M e Y. Per ciascuna variabile, si ottiene un punteggio previsto utilizzando il punteggio al tempo 1 della stessa variabile come predittore. Successivamente, si procede con

l'analisi utilizzando i punteggi residui di X, M e Y, che rappresentano il cambiamento di queste variabili con il punteggio al tempo 1 rimosso. Nel caso della variabile dipendente, si stima un unico effetto mediato che lega il cambiamento corretto di X al cambiamento corretto di M, il quale, a sua volta, è associato al cambiamento corretto di Y.
L'analisi dei punteggi di cambiamento corretti permette di ottenere una comprensione più approfondita dei processi di mediazione e delle relazioni tra le variabili nel corso del tempo, fornendo una prospettiva dettagliata su come le variabili si influenzino reciprocamente all'interno del contesto longitudinale.

Un'ulteriore alternativa consiste nell'utilizzare le misurazioni al tempo 1 come covariate in un modello di analisi di covarianza, come mostrato nelle equazioni (3.5) e (3.6):

$$Y_2 = \beta_1 + c_1' X_1 + c_2' X_2 + b_1 M_1 + b_2 M_2 + s_1 Y_1 + e_1, \tag{3.5}$$

$$M_2 = \beta_2 + a_1 X_1 + a_2 X_2 + s_2 M_1 + e_2. \tag{3.6}$$

Come nelle equazioni delle sezioni precedenti e_1 ed e_2 rappresentano gli errori casuali, mentre β_1 e β_2 le intercette. Questo modello viene denominato condizionale in quanto i punteggi alla seconda misurazione dipendono dai punteggi della prima misurazione. L'equazione (3.5) rappresenta la relazione di X_1, X_2, M_1, M_2 e Y_1 con Y_2, dove il coefficiente s_1 riflette la stabilità temporale della variabile Y, dopo l'aggiustamento per le altre variabili. Vi sono diversi possibili stimatori dell'effetto di mediazione: $\hat{a}_1\hat{b}_1$ rappresenta le relazioni nel tempo, mentre $\hat{a}_2\hat{b}_2$ riflette le relazioni tra le osservazioni alla seconda misurazione. Lo stimatore $\hat{a}_1\hat{b}_1$ risulta essere una misura migliore dell'effetto mediato, poiché $\hat{a}_1$ rappresenta la relazione temporale tra X_1 e M_2 e $\hat{b}_1$ rappresenta quella tra M_1 e Y_2 (Cole e Maxwell, 2003). Si possono applicare le formule degli *standard error* e degli intervalli di confidenza descritti nelle sezioni precedenti per valutare la significatività degli effetti di mediazione.

Una quarta soluzione consiste nell'utilizzo di un modello di mediazione autoregressivo, in cui ogni variabile viene predetta dalla stessa variabile alla misurazione precedente. In questo modello, non vengono incluse relazioni contemporanee tra le variabili, ovvero i parametri c_2', b_2 e a_2 delle equazioni (3.5) e (3.6). Come nella prima misurazione, gli

errori nelle variabili alla seconda misurazione possono covariare, riflettendo la possibilità di relazioni contemporanee tra le variabili, ma senza specificare la loro direzione.
L'uso di questo modello permette di considerare la struttura dinamica dei dati longitudinali, tenendo conto delle interazioni temporali e delle relazioni reciproche tra le variabili nel tempo.

Ognuno di questi modelli adotta un approccio distintivo per rappresentare i cambiamenti nel tempo. I modelli che si basano sulla differenza tra i punteggi assumono che le differenze iniziali rimangano costanti nelle misurazioni successive. L'analisi di covarianza e l'analisi dei punteggi di cambiamento corretti assumono invece che i punteggi tendano a convergere verso la media nel corso del tempo.

Le qualità e le limitazioni insite in questi modelli suggeriscono un approccio generale all'analisi di mediazione in contesto longitudinale. In primo luogo il modello di crescita latente offre un quadro generale per valutare l'eventuale variazione nel tempo e come questa possa cambiare fra i soggetti. Il procedimento inizia con la costruzione di modelli di crescita individuali per le singole variabili. Se emerge una significativa crescita temporale nelle variabili, si può ipotizzare l'adeguatezza di un modello di crescita latente. Dopo aver conseguito modelli soddisfacenti per ciascuna variabile, si passa alla creazione di un modello combinato che considera le relazioni intercorrenti tra X, M e Y. Qualora non si rilevi una variazione sostanziale nel tempo, il modello autoregressivo diventa spesso l'opzione prediletta, data la sua capacità di esaminare le relazioni di mediazione in contesti sia longitudinali che trasversali. Questo modello è inoltre in grado di incorporare configurazioni più intricate, tra cui la mediazione con percorsi multipli.

3.3 Modello multilivello

I modelli multilivello sono specificamente orientati all'analisi di dati che presentano una struttura gerarchica o a *cluster*. Tali dati si riscontrano comunemente in diversi campi, ad esempio nella ricerca educativa con gli studenti inseriti nelle scuole, negli studi familiari con i bambini inseriti nelle famiglie o nella ricerca medica con i pazienti inseriti negli ospedali. I dati suddivisi in *cluster* possono anche derivare da un disegno di ricerca specifico. Un esempio sono i disegni longitudinali, dove i dati sono considerati come una serie di misurazioni ripetute all'interno dei singoli soggetti.

L'introduzione dei gruppi comporta un ulteriore livello di complessità che richiede una considerazione adeguata nell'analisi dei dati. I risultati di due osservazioni appartenenti allo stesso *cluster* tendono spesso ad essere più simili tra loro rispetto ai risultati di due osservazioni provenienti da *cluster* diversi, anche dopo aver tenuto conto delle caratteristiche dei soggetti. Questa omogeneità interna ai gruppi viola l'assunzione sottostante la maggior parte dei modelli di regressione, che presuppone l'indipendenza delle osservazioni.

Un concetto chiave nell'analisi multilivello è il coefficiente di correlazione intragruppo (ICC). L'ICC (3.7) quantifica il grado di somiglianza tra gli individui appartenenti allo stesso gruppo. Da una prospettiva epidemiologica, questo consente di stabilire in che misura la variabilità dei risultati possa essere attribuita all'unità di raggruppamento. Ad esempio, si può stabilire quanto della variazione nei risultati dei pazienti sia riconducibile all'ospedale presso cui sono stati ricoverati. L'indice è dato da:

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}, \tag{3.7}$$

dove MS_B rappresenta l'errore quadratico medio tra i gruppi, MS_W l'errore quadratico medio interno ai gruppi e k denota il numero di soggetti presenti in ciascun gruppo. Il valore dell'ICC varia nell'intervallo tra $\frac{-1}{k-1}$ e 1.

Qualora il valore dell'ICC differisse da zero, l'analisi di mediazione verrebbe affetta da una violazione dell'assunzione di indipendenza. In particolare, se i soggetti tendono a manifestare risposte simili, l'ICC assumerà un valore positivo.
La significatività del coefficiente di correlazione intragruppo può essere verificata mediante la seguente statistica:

$$F_{g-1,g(k-1)} = \frac{1 + (k-1)ICC}{1 - ICC}, \tag{3.8}$$

dove g rappresenta il numero di gruppi e k il numero di individui in ciascun gruppo.

Nonostante effettuare il *test* per la significatività dell'ICC costituisca un primo passo importante per determinare se viene violata l'assunzione di indipendenza in un campione di dati, è importante notare che anche valori dell'ICC molto modesti possono influenzare i *test* di significatività, specialmente se il numero di individui in ogni gruppo è considerevole (Kreft, 1996; Barcikowski, 1981).

Mediante l'impiego di un metodo di regressione tradizionale, quando l'assunzione di indipendenza viene violata, le stime dei coefficienti di regressione e dei relativi errori *standard* possono subire distorsioni. Trattare le variabili a livello di gruppo come se fossero misurate a livello individuale può comportare una sottostima degli errori *standard*, con la conseguente produzione di risultati erroneamente significativi e intervalli di confidenza artificialmente ristretti.

È pertanto necessario adottare un modello di analisi multilivello, che risulta più complesso rispetto al modello di mediazione singola, e solitamente richiede un approccio iterativo per la stima dei parametri. La variabile indipendente opera a livello di gruppo, mentre sia il mediatore che la variabile dipendente agiscono a livello individuale.
Al primo livello viene specificato un modello per gli individui all'interno di ciascun gruppo, in cui i parametri vengono considerati casuali e variano in parte in relazione ai predittori a livello di gruppo. Al secondo livello viene costruito un ulteriore modello lineare, ma in questa circostanza, l'intercetta del primo livello diventa la variabile dipendente del secondo livello. Di conseguenza, si hanno formulazioni a livello individuale e a livello di gruppo per ciascuna delle tre equazioni descritte nel modello singolo (3.1, 3.2, 3.3).

Le equazioni per Y predetto da X sono:

$$\text{Livello individuale 1: } Y_{ij} = \beta_{0j} + e_{ij}, \tag{3.9}$$

$$\text{Livello di gruppo 2: } \beta_{0j} = \gamma_{00} + cX_j + u_{0j}. \tag{3.10}$$

In queste equazioni l'indice i si riferisce all'individuo e j al gruppo. Si osserva che il valore della variabile dipendente nell'equazione (3.9) comprende anche un errore casuale a livello individuale e_{ij}, che si assume segua una distribuzione normale. Nell'equazione (3.10) γ_{00} rappresenta la media complessiva, c il coefficiente angolare e u_{0j} costituisce la deviazione casuale tra la media prevista e quella osservata a livello di gruppo, che si assume avere distribuzione normale.

Le equazioni per Y predetto da X e M sono:

$$\text{Livello individuale 1: } Y_{ij} = \beta_{0j} + bM_{ij} + e_{ij}, \tag{3.11}$$

$$\text{Livello di gruppo 2: } \beta_{0j} = \gamma_{00} + c'X_j + u_{0j}. \tag{3.12}$$

Queste equazioni comprendono due predittori: M a livello individuale e X a livello di gruppo. Si possono creare anche altri modelli che includono X e M come predittori.

Le equazioni per M predetto da X sono:

$$\text{Livello individuale 1: } M_{ij} = \beta_{0j} + e_{ij}, \tag{3.13}$$

$$\text{Livello di gruppo 2: } \beta_{0j} = \gamma_{00} + aX_j + u_{0j}. \tag{3.14}$$

Data questa serie di equazioni, emerge che il parametro b è l'unico stimato a livello individuale, mentre i parametri c, c' e a vengono stimati a livello di gruppo.

Il modello multilivello presenta una struttura di notevole complessità, e i suoi parametri non possono essere stimati mediante formule esatte, ma richiedono l'utilizzo di approcci iterativi, quali le tecniche di massima verosimiglianza vincolata (REML).

Quando si analizzano congiuntamente gli effetti del predittore sia a livello di gruppo che a livello individuale, diviene rilevante applicare la centratura, che generalmente comporta la rimozione della media di gruppo dai predittori. Se una variabile X non è stata centrata, l'intercetta rappresenta il valore di Y quando X è pari a zero, anche se il valore zero per X risulta inverosimile. Mediante la centratura, l'intercetta diviene invece il valore di Y corrispondente al valore medio di X. Vi sono tre principali metodologie per riscalare i predittori all'equazione di livello 1 (Hofmann e Gavin, 1998): la misura grezza, in cui non avviene nessuna centratura e le variabili del primo livello mantengono la loro scala originale; la centratura rispetto alla media globale, che prevede la sottrazione della media complessiva dai predittori; e la centratura rispetto alla media di gruppo, in cui la media di ogni gruppo viene sottratta dai valori osservati per ciascun soggetto all'interno del gruppo.

Capitolo 4

Applicazione su dataset

Le linee guida del *National Institute of Health* (2007) definiscono l'asma come segue: "L'asma è un disturbo infiammatorio cronico delle vie aeree in cui molti tipi di cellule ed elementi cellulari svolgono un ruolo: in particolare, cellule dei mastociti, eosinofili, linfociti T, macrofagi, neutrofili e cellule epiteliali. In individui suscettibili, questa infiammazione causa episodi ricorrenti di sibilanza, dispnea, oppressione al petto e tosse, in particolare di notte o al mattino presto. Questi episodi sono generalmente associati a ostruzione delle vie aeree diffusa ma variabile, spesso reversibile spontaneamente o con trattamento".

In questo capitolo, si presenta l'applicazione della teoria di analisi di mediazione continua a un *dataset* che coinvolge pazienti affetti da asma. Dopo aver delineato i fondamenti teorici precedentemente esposti, si affronta ora il processo di applicazione della metodologia a dati reali, allo scopo di comprendere meglio le dinamiche sottostanti l'asma e i suoi fattori associati. Il dataset è costituito da 100 soggetti asmatici, che vengono valutati prima e dopo il *test* di broncodilatazione con salbutamolo, utilizzato per valutare la funzione polmonare e la risposta delle vie aeree al farmaco.

I risultati riportati di seguito sono il prodotto di un'analisi condotta utilizzando il software R (http://www.r-project.org/) nella versione 4.2.3. Il livello di significatività è fissato al 5%. Per i principali codici impiegati nell'esecuzione delle analisi e per ulteriori dettagli sul modello, si fa riferimento all'Appendice (A, B).

4.1 Descrizione dei dati

Al fine di valutare il modello di mediazione singola, le variabili considerate nell'applicazione sono le seguenti:

- **feNO**: variabile quantitativa che esprime la differenza fra la concentrazione di ossido nitrico nell'aria esalata dai polmoni misurata dopo il *test* di broncodilatazione con salbutamolo e quella misurata prima del *test*.
- **GINA**: variabile dicotomica che rappresenta il sistema di valutazione utilizzato dalla *Global Initiative for Asthma* (2023) per classificare il controllo dell'asma in base ai sintomi e alla gestione della malattia. In particolare le due modalità sono: `W` (*Well-controlled*), modalità di riferimento, e `PU` (*Partially controlled* e *Uncontrolled*).
- **R5R20**: variabile quantitativa che misura la funzione polmonare valutando la resistenza delle vie aeree in risposta a onde sonore di diverse frequenze, 5 Hz e 20 Hz.
- **sesso**: variabile dicotomica che indica il sesso del paziente, ovvero maschio (sesso = `1`) o femmina (sesso = `0`).
- **età**: variabile quantitativa che indica l'età del paziente.
- **durata**: variabile quantitativa che indica la durata in anni dell'asma per ogni paziente.
- **gruppo**: variabile dicotomica che indica se, a seguito del *test* di broncodilatazione, la concentrazione di ossido nitrico esalata dal paziente è aumentata (gruppo = `1`) o diminuita (gruppo = `0`).
- **fumo**: variabile dicotomica che indica se il paziente fuma (fumo = `1`) o no (fumo = `0`).
- **BMI**: variabile quantitativa che rappresenta l'indice di massa corporea per ogni paziente.

In fase di analisi sono state escluse le covariate *fumo*, *durata* e *sesso*, poiché i loro effetti stimati sono risultati non significativi in tutte e tre le equazioni di regressione (Tabelle B.1, B.2, B.3).

4.2 Modello di mediazione

Il modello di mediazione singola esaminato in questo capitolo utilizza la variabile *feNO* come variabile dipendente Y, *GINA* come variabile indipendente X e *R5R20* come mediatore M. Le restanti variabili rappresentano le covariate del modello.

L'obiettivo di questo studio è esaminare come il livello di gravità dell'asma del paziente (*GINA*) influisca sull'efficienza del *test* di broncodilatazione con salbutamolo (*feNO*), e come questo effetto venga mediato dalla funzione polmonare del paziente (*R5R20*).

Il modello di mediazione risultante dall'applicazione al *dataset* è rappresentato dalle seguenti equazioni:

$$\hat{Y} = 6.80 - 5.91X + 0.110E + 16.4G - 0.356B, \tag{4.1}$$

$$\hat{M} = -0.0137 - 0.0821X + 0.00113E + 0.0665G + 0.00363B, \tag{4.2}$$

$$\hat{Y} = 7.29 - 2.93X + 36.4M + 0.0686E + 13.9G - 0.488B, \tag{4.3}$$

dove E rappresenta la variabile *età*, G la variabile *gruppo* e B la variabile *BMI*.
L'ipotesi di esistenza di un'interazione tra il mediatore e la variabile indipendente viene respinta (*p-value* $= 0.676$).

Nelle equazioni (4.1) e (4.3), i valori negativi della variabile indipendente X riflettono come i pazienti con un livello di controllo dell'asma più elevato (*GINA* $=$ *Well-controlled*), tendano a registrare valori più bassi della variabile di risposta Y. Ciò implica che l'efficacia del *test* di broncodilatazione con salbutamolo varia in base alla gravità della malattia, con una riduzione significativa nei pazienti che manifestano sintomi meno gravi.

Si osserva inoltre che il coefficiente $\hat{b}$ è pari a 36.4, il che suggerisce che, mantenendo costanti le altre variabili, un aumento nella variabile *R5R20* è associato a una maggiore differenza tra la concentrazione di ossido nitrico esalata dai polmoni prima e dopo il *test*.

Esaminando le altre covariate, emerge che un aumento dell'età del paziente o un incremento nella concentrazione di ossido nitrico esalata in seguito al *test* di broncodilatazione sono associati a valori maggiori di *feNO*. Questo suggerisce che sia in presenza

che in assenza del mediatore, le variabili *età* e *gruppo* influenzano l'efficacia del *test* di broncodilatazione con salbutamolo. D'altra parte, un incremento nell'indice di massa corporea si associa a una riduzione nella variabile di risposta.

Nel quadro di questo modello, l'effetto totale è pari a $\hat{c}$ = -5.91, l'effetto diretto a $\hat{c}$' = -2.93 e l'effetto indiretto, o effetto mediato, a $\hat{a}\hat{b} = \hat{c} - \hat{c}$' = -2.98, con intervallo di confidenza pari a (-6.16, -0.67) (Tabella 4.1).

	Stima	**IC**	**p-value**
Effetto mediato	-2.98	(-6.16, -0.67)	0.004
Effetto diretto	-2.93	(-7.65, 1.50)	0.224
Effetto totale	-5.91	(-10.46, -1.22)	0.012

Tabella 4.1: Stime, intervalli di confidenza e *p-value* degli effetti stimati del modello di mediazione.

La proporzione dell'effetto totale che viene mediata è pari a $\frac{\hat{a}\hat{b}}{\hat{a}\hat{b}+\hat{c}'}$. = 0.505. Questo risultato implica che approssimativamente il 50% dell'effetto totale che esercita il livello di gravità dell'asma sulla variazione della concentrazione di ossido nitrico esalato dal paziente a seguito del *test* viene mediato dalla funzione polmonare dello stesso.

Si procede ora con la verifica della mediazione utilizzando il metodo degli *step* causali descritto da Baron e Kenny (1986), precedentemente presentato nella Sezione 2.2. Il primo passaggio consiste nel determinare se la variabile indipendente *GINA* influisce sulla variabile dipendente *feNO*. Con un *p-value* pari a 0.0468, il coefficiente stimato $\hat{c}$ risulta essere statisticamente significativo, confermando in questo modo la presenza di un effetto da mediare. Si prosegue quindi testando l'effetto della variabile X sul mediatore: il coefficiente $\hat{a}$ è significativo (*p-value* = 0.00141), il che indica l'esistenza di una relazione tra la variabile *GINA* e la variabile *R5R20*. Inoltre, il mediatore deve avere un effetto sulla variabile dipendente Y quando la variabile X è controllata, ovvero il coefficiente $\hat{b}$ deve risultare significativo. Con un *p-value* pari a 0.00343 si conferma la relazione fra la variabile *R5R20* e la variabile *feNO*. Infine, viene esaminato l'effetto diretto. Il coefficiente $\hat{c}$' risulta essere non statisticamente significativo, con *p-value* = 0.328. Pertanto, si tratta di un caso di mediazione totale, in cui la funzione polmonare del paziente media interamente la relazione fra il livello di gravità dell'asma e la va-

riazione della concentrazione di ossido nitrico esalato dai polmoni a seguito del *test* di broncodilatazione con salbutamolo.

4.3 Analisi di sensibilità

L'analisi di sensibilità è una procedura impiegata nell'analisi di mediazione statistica, con l'obiettivo di valutare quanto i risultati siano influenzati da variazioni nelle specifiche del modello o in altre assunzioni cruciali. Tale analisi fornisce un contributo informativo di rilievo per valutare la robustezza delle conclusioni dell'analisi di mediazione.

Gli esiti dell'analisi di sensibilità svolta sul *dataset* considerato (Tabella 4.2) mostrano come l'effetto di mediazione sia influenzato da differenti valori di ρ (coefficiente di correlazione tra M e Y), confermando la relazione stretta tra *R5R20* e *feNO*.

ρ	**Effetto mediato**	**IC**
0.1	-2.06	(-4.30, 0.177)
0.2	-1.11	(-3.10, 0.889)
0.3	-0.0925	(-1.98, 180)
0.4	1.03	(-0.953, 3.01)

Tabella 4.2: Analisi di sensibilità.

Nella Figura (4.1) la linea orizzontale tratteggiata rappresenta l'effetto di mediazione stimato sotto il modello, mentre la linea continua rappresenta l'effetto di mediazione al variare di diversi valori di ρ. La regione grigia rappresenta invece gli intervalli di confidenza al 95%.

Si può notare che la conclusione sulla direzione dell'effetto di mediazione ($\hat{a}\hat{b}$ = -2.98) rimane invariata finchè ρ risulta inferiore a 0.1, punto in cui gli intervalli di confidenza iniziano a includere il valore zero. Il valore esatto per cui l'effetto di mediazione diventa nullo è $\rho = 0.3$.

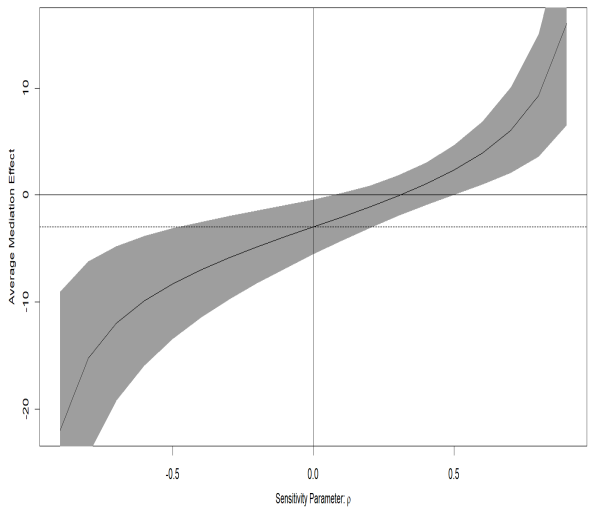

Figura 4.1: Rappresentazione grafica dei risultati dell'analisi di sensibilità.

In conclusione, poichè l'effetto di mediazione risulta influenzato dai differenti valori della correlazione fra M e Y, potrebbero esistere delle covariate pre o post *test* non considerate che confondono la relazione tra il mediatore e la risposta.

Capitolo 5

Conclusioni

Il presente studio si è posto l'obiettivo di approfondire il ruolo del mediatore, con il fine di offrire un'adeguata comprensione del suo utilizzo in diversi contesti di analisi statistica.

Sono emersi in particolare gli aggiustamenti necessari quando si utilizza un modello di mediazione con la variabile dipendente dicotomica. In tali modelli, è essenziale impiegare una variabile latente sottostante come variabile dipendente. Fra i metodi presentati, quelli basati su $\hat{a}\hat{b}$ risultano essere i più accurati, in quanto non suscettibili ai problemi di scala che si manifestano quando si utilizza una variabile latente continua. Quando ci si confronta invece con più mediatori, si incorporano effetti specifici di mediazione per ciascuno di essi, e l'effetto totale di mediazione è il risultato della somma di tutti questi effetti specifici.

Il modello di mediazione singola continua è caratterizzato dall'utilizzo equivalente degli stimatori $\hat{a}\hat{b}$ e $\hat{c} - \hat{c}$' per il calcolo dell'effetto di mediazione. Un elemento di rilievo di questo tipo di analisi consiste nella verifica dell'effetto di mediazione, che può essere eseguita attraverso varie metodologie, di cui la più utilizzata è rappresentata dal metodo degli *step* causali di Baron e Kenny (1986), il quale consente anche di stabilire se si tratti di una mediazione parziale o completa. Estensioni del modello di mediazione singola includono il modello longitudinale e il modello multilivello. L'analisi del modello longitudinale ha rivelato alcuni limiti derivanti dall'utilizzo improprio di dati trasversali per l'analisi di misure ripetute; ha inoltre evidenziato i tre concetti fondamentali di stabilità, stazionarietà ed equilibrio. Di conseguenza, sono state esplorate diverse metodologie atte ad affrontare questa tipologia di analisi, in particolare in relazione ai modelli di regressione a due ondate. Infine, è stato affrontato il modello multilivello, associa-

to all'importante concetto del coefficiente di correlazione intragruppo, il quale misura il grado di violazione dell'assunzione di indipendenza tra le osservazioni all'interno dei gruppi.

L'applicazione al *dataset* relativo all'asma ha consentito di testare il modello di mediazione singola attraverso un'analisi di dati reali. Da questa analisi è emerso come l'efficacia del test di broncodilatazione con salbutamolo risulti ridotta nei pazienti che presentano sintomi meno gravi della patologia. Esaminando il ruolo del mediatore, si è stabilito come circa il 50% dell'effetto totale esercitato dal livello di gravità dell'asma sulla variazione della concentrazione di ossido nitrico esalato dal paziente in risposta al *test* viene mediato dalla sua funzione polmonare. L'effetto di mediazione è risultato significativo e si è confermato che si tratta di un caso di mediazione totale, in cui la relazione fra X e Y risulta significativa solo se mediata da M. L'analisi di sensibilità ha convalidato la relazione fra M e Y e ha determinato che il coefficiente di correlazione, per il quale l'effetto di mediazione si annulla, è pari a 0.3.

L'analisi di mediazione si rivela come uno strumento cruciale per decifrare i meccanismi sottili che guidano le relazioni tra le variabili. Questa tesi vuole mostrare come, dietro l'apparente semplicità del mediatore, si nasconda una rete intricata di cause ed effetti, la cui analisi consente di effettuare valutazioni in modo più informato e consapevole.

Appendice A

Codici R

I codici più importanti utilizzati per svolgere l'analisi di mediazione sul *dataset* sull'asma sono:

```
library(mediation)
y.mod1 <- lm(feNO ~ GINA + eta + gruppo + BMI, data=asma)
m.mod <- lm(R5R20 ~ GINA + eta + gruppo + BMI, data=asma)
y.mod2 <- lm(feNO ~ R5R20 + GINA + eta + gruppo + BMI, data=asma)
med.out <- mediate(m.mod, y.mod2, treat="GINA", mediator = "R5R20",
boot=TRUE, sims=500)
```

L'interazione fra la variabile X e la variabile M è stata esaminata attraverso le seguenti funzioni della libreria "*mediation*":

```
y.mod2.int <- lm(feNO ~ R5R20 * GINA + eta + gruppo + BMI, data=asma)
med.out.int <- mediate(m.mod, y.mod2.int, treat="GINA", mediator = "R5R20",
boot=TRUE, sims=500)
test.TMint(med.out.int, conf.level = .95)
```

L'analisi della sensibilità viene svolta con il codice che segue:

```
sens <- medsens(med.out, rho.by = 0.1)
plot(sens, sens.par = "rho", main="")
```

Appendice B

Dataset completo

Le stime dei coefficienti delle tre equazioni di regressione per il modello che incorpora tutte le variabili presenti nel *dataset* iniziale, sono illustrate nelle tabelle che seguono.

	Stima	**Std. Error**	$\mathbf{t}_{oss}$	**p-value**
$\beta_{intercetta}$	6.74	5.91	1.14	0.257
β_{GINA}	-5.97	2.98	-2.00	0.0484
β_{sesso}	-0.437	2.46	-0.178	0.859
$\beta_{età}$	0.119	0.0776	1.54	0.129
β_{durata}	-0.0308	0.119	-0.259	0.796
β_{gruppo}	16.1	2.54	6.33	<0.001
β_{fumo}	-1.19	2.57	-0.462	0.646
β_{BMI}	-0.322	0.229	-1.41	0.163

Tabella B.1: Coefficienti stimati per l'equazione 3.1.

	Stima	**Std. Error**	$\mathbf{t}_{oss}$	**p-value**
$\beta_{intercetta}$	-0.0173	0.0484	-0.357	0.722
β_{GINA}	-0.0829	0.0244	-3.39	0.00105
β_{sesso}	-0.0306	0.0202	-1.52	0.133
$\beta_{età}$	0.00129	0.000637	2.03	0.0459
β_{durata}	-0.000432	0.000975	-0.443	0.659
β_{gruppo}	0.0579	0.0208	2.78	0.00676
β_{fumo}	-0.0286	0.0211	-1.36	0.178
β_{BMI}	0.00495	0.00188	2.64	0.00995

Tabella B.2: Coefficienti stimati per l'equazione 3.2.

	Stima	**Std. Error**	$\mathbf{t}_{oss}$	**p-value**
$\beta_{intercetta}$	7.38	5.67	1.30	0.196
β_{R5R20}	36.9	12.8	2.89	0.00485
β_{GINA}	-2.91	3.05	-0.954	0.343
β_{sesso}	0.694	2.39	0.290	0.772
$\beta_{età}$	0.0716	0.0762	0.938	0.351
β_{durata}	-0.0148	0.114	-0.130	0.897
β_{gruppo}	13.9	2.55	5.47	$<$0.001
β_{fumo}	-0.129	2.49	-0.052	0.959
β_{BMI}	-0.505	0.228	-2.21	0.0298

Tabella B.3: Coefficienti stimati per l'equazione 3.3.

Ringraziamenti

Desidero in primo luogo esprimere il mio sincero ringraziamento alla Professoressa Laura Ventura, che, oltre ad avermi seguita con grande disponibilità durante la stesura della tesi, mi ha introdotta al mondo della statistica medica, un campo che ha suscitato in me un interesse profondo e un apprezzamento senza precedenti.

Vorrei poi dedicare questa tesi a tutti coloro i quali hanno contribuito a plasmare la persona che sono diventata oggi. Primi fra tutti i miei genitori, che, con l'amore smisurato di chi desidera solo il meglio per te, da sempre sono stati un'ancora nei momenti di difficoltà e mi hanno insegnato che bisogna lottare per raggiungere grandi soddisfazioni. Ringrazio quindi mio fratello, che mi ha mostrato in prima persona cosa vuol dire puntare alle stelle e che rappresenta per me un esempio di ingegno e tenacia. Allo stesso modo ringrazio i miei nonni, anche quelli che non ci sono più, per avermi sostenuta e amata incondizionatamente fin da quando sono piccola.
Un grande apprezzamento va anche a mia cugina, per essere stata un'amica, una guida e una sicurezza durante tutta la mia vita; ai miei amici più cari, Aurora, Nikol, Emanuele e Nicolò, per essere stati al mio fianco e aver sempre tifato per me durante gli anni delle superiori e dell'università. Un ringraziamento va inoltre agli zii, a tutti gli altri amici e parenti e ad ogni persona che ho incontrato durante il mio percorso di vita, e che in piccolo o in grande mi ha plasmata nella persona che sono ora.

Infine non posso che ringraziare me stessa, per non aver mai mollato, anche quando sono stata messa a dura prova o quando mi sembrava di fare sacrifici senza ricevere soddisfazioni. Nell'università ho ritrovato me stessa, con tutta la passione e la forza di chi è consapevole delle potenzialità che possiede.

Sara Bisson

Bibliografia

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Stat Med*, *36*(20), 3257–3277.

Barcikowski, R. S. (1981). Statistical Power with Group Mean as the Unit of Analysis. *Journal of Educational Statistics*, *6*(3), 267–285.

Baron & Kenny. (1986). *Steps to Establish Mediation*.

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Bayman, E. O., & Dexter, F. (2021). Multicollinearity in Logistic Regression Models. *Anesthesia & Analgesia*, *133*(2), 362–365.

Burr, J. A., & Nesselroade, J. R. (1990). Change Measurement. In A. von Eye (Cur.), *Statistical Methods in Longitudinal Research* (pp. 3–34). Academic Press.

Cole, D. A., & Maxwell, S. E. (2003). Testing Mediational Models With Longitudinal Data: Questions and Tips in the Use of Structural Equation Modeling. *Journal of Abnormal Psychology*, *112*(4), 558–577.

De la Cruz, R., Marshall, G., & Quintana, F. A. (2011). Logistic regression when covariates are random effects from a non-linear mixed model. *Biometrical Journal*, *53*(5), 735–749.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. Oxford.

Freedman, S. (1992). Sample size for studying intermediate endpoints within intervention trials of observational studies. *American Journal of Epidemiology*, *136*(9), 1148–1159.

Fritz, M. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233–239.

Global Initiative for Asthma. (2023). Global Strategy for Asthma Management and Prevention - 2023.

Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and Estimating Indirect Effects Assuming Time Lags Really Matter. In L. M. Collins & J. L. Horn (Cur.), *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions* (pp. 243–259). American Psychological Association.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*(5), 623–641.

James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods*, *9*(2), 233–244.

Judd, C., & Kenny, D. (1981). *Estimating the Effects of Social Intervention*. Cambridge University Press.

Kenny, D. (1979). *Correlation and Causality* (Vol. -1).

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2S), S101–S108.

Kreft, I. G. G. (1996). Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies [Retrieved Month XX, 200X].

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*, 281–299.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis Group.

MacKinnon, D., & Dwyer, J. (1993). Estimating Mediated Effects in Prevention Studies. *Evaluation Review*, *17*, 144–158.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & Wang, E. Y. (1991). Mediating mechanisms in a school-based drug prevention program: first-year effects of the Midwestern Prevention Project. *Health Psychology*, *10*(3), 164–172.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical trials*, *4*(5), 499–513.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

National Heart, Lung, and Blood Institute. (2007). Guidelines for the Diagnosis and Management of Asthma (EPR-3) [Accessed July 16, 2015].

Perneger, T., Combescure, C., & Poncet, A. (2023). Adjustment for baseline characteristics in randomized trials using logistic regression: sample-based model versus true model. *Trials*, *24*(1), 107.

Shultz, T. (1982). Rules of Causal Attribution. *Monographs of the Society for Research in Child Development*, *47*(1), Serial No, 194.

Stoltzfus. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099–1104.

VanderWeele. (2016). Explanation in causal inference: developments in mediation and interaction. *Int J Epidemiol.*, *45*(6), 1904–1908.

Winship, C., & Mare, R. D. (1983). Structural Equations and Path Analysis for Discrete Datal. *The American Journal of Sociology*, *89*(1), 54–110.