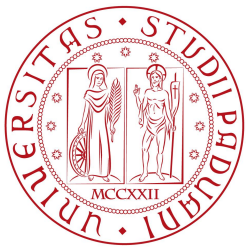
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche



Eterogeneità non osservata e identificazione di effetti causali mediante modelli mistura

Relatrice: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Correlatore: prof. Marco Bertoni
Dipartimento di Scienze Economiche e Aziendali "Marco Fanno"

Laureando: Leonardo Genesin
Matricola n. 2056404

Anno Accademico 2022/2023

Indice

Introduzione

Uno degli obiettivi principali dell'analisi statistica è lo studio delle relazioni tra variabili e a questo scopo lo strumento principe è rappresentato dai modelli di regressione. Un errore comune d'interpretazione dei risultati che si ottengono da un modello di regressione è quello di attribuire un rapporto di causa-effetto tra le variabili esplicative e la variabile risposta. Sebbene questo tipo di interpretazione possa talvolta risultare ragionevole, è necessario tener presente che un modello di regressione è costruito, dal punto di vista statistico, per trarre conclusioni di natura associazionale, non causale, tra le variabili.

Oggetto di questa tesi è lo studio dei metodi, noti sotto il cappello di *inferenza causale*, che permettono, invece, di rispondere a quesiti di natura causale. Viene fatto largo uso di metodi di inferenza causale nei contesti socio-economici, dove è spesso d'interesse conoscere l'effetto che un determinato trattamento o una politica d'intervento ha su una variabile d'interesse. In tali contesti, infatti, si trattano il più delle volte dati proventienti da studi osservazionali, dove le variabili sono osservate sugli individui senza alcun controllo o disegno sperimentale precedente alla raccolta dei dati. La principale sfida da affrontare, in questo caso, è quella di isolare l'effetto che il trattamento ha sulla variabile d'interesse da possibili altre fonti che possono avere un'influenza su di essa. Non essendoci un controllo diretto sugli individui della popolazione che si sottopongono o non si sottopongono al trattamento, per poterne isolare l'effetto è necessario tener conto del fatto che esistono ragionevolmente delle componenti di eterogeneità nei due distinti gruppi di individui, che prescindono dal trattamento oggetto di studio e che possono influire sia sulla loro decisione a sottoporsi al trattamento sia sulla loro risposta al trattamento. Ignorare tali fonti di eterogeneità nella valutazione dell'effetto del trattamento significa indurre distorsione nella stima. Tra le diverse tecniche proposte a soluzione di questo problema, una delle più semplici prevede di controllare per l'eterogeneità negli individui, includendo più caratteristiche individuali possibili tra quelle che si pensa possano interagire con il trattamento in analisi, in modo da ottenere una quantificazione dell'effetto causale d'interesse più corretta.

Un ulteriore oggetto di studio in questa tesi è rappresentato dai modelli mistura, una classe di modelli statistici molto flessibili che permette di rappresentare e analizzare dati complessi. L'utilizzo dei modelli mistura è motivato in con-

testi in cui è ragionevole assumere che il fenomeno di interesse si manifesti in modo eterogeneo all'interno della popolazione, ad esempio perché riconducibile alla presenza di sottopopolazioni definite a partire da una o più caratteristiche non osservate, o non osservabili, e per questo dette latenti. L'inferenza su un modello mistura, quando si utilizza un approccio parametrico, è spesso basato su procedure legate alla teoria della verosimiglianza e consente, a partire dalla stima dei parametri, l'identificazione delle classi latenti di eterogeneità.

L'obiettivo di questa tesi è quello di esplorare la possibilità di identificare eventuali fonti di eterogeneità latente, mediante il ricorso a un modello di regressione a mistura finita, volto a valutare l'effetto di un trattamento. A questo scopo, si propone una metodologia che si configura come un modello a più equazioni, che simultaneamente descrivono, entro ciascuna classe latente di eterogeneità, la relazione tra la variabile risposta e il trattamento, al netto di un insieme di caratteristiche di controllo, e gestiscono l'endogeneità del trattamento. La metodologia proposta ha il vantaggio di essere uno strumento flessibile che può essere applicato a diversi contesti di analisi e dati di arbitraria natura, e permette di ottenere risultati articolati, individuando e valutando anche effetti del trattamento eterogenei per diverse tipologie di individui.

Si presenta inoltre un algoritmo EM appositamente adattato, per la determinazione della stima di massima verosimiglianza dei parametri del modello. L'efficacia della metodologia viene valutata mediante uno studio di simulazione ad ampio spettro ed un'applicazione a dati reali.

La trattazione si sviluppa come segue. Nel Capitolo 1 viene presentato il problema della valutazione dell'effetto di un trattamento, quali criticità presenta e come queste vengono affrontate con i metodi tradizionali dell'inferenza causale, con un particolare focus sull'utilizzo di metodi di regressione. Nel Capitolo 2 vengono introdotti i modelli di regressione a mistura finita e discussi gli aspetti di stima dei parametri, di valutazione della loro incertezza e di selezione del modello. Nel Capitolo 3 viene proposto un modello mistura di regressione a due livelli che può essere utilizzato nell'ambito di valutazione dell'effetto di un trattamento come strumento per affrontare il problema dell'eterogeneità non osservata. Vengono presentati gli aspetti inferenziali del modello e ne vengono discussi vantaggi e punti critici. Nel Capitolo 4 si presentano i risultati di uno studio di simulazione volto a valutare l'efficacia del modello proposto al variare di diverse fonti di variabilità. Nel Capitolo 5 viene infine presentata un'applicazione del modello proposto ad un insieme di dati reali, con l'obiettivo di valutare l'effetto che l'istruzione ha sul reddito, confrontandone i risultati con le analisi svolte in precedenza con altri metodi. Si presentano infine alcune considerazioni conclusive.

Capitolo 1

La valutazione dell'effetto di un trattamento

L'inferenza causale è un ambito della statistica che necessita di una propria formalizzazione che permetta di rispondere adeguatamente ad un quesito di natura causale d'interesse. In questo capitolo si introducono il contesto nonché la notazione di riferimento, diversa rispetto a quella tipicamente usata nell'analisi statistica classica, e si presentano i problemi e le assunzioni tipici di un'analisi di tipo causale, prima di procedere con una breve rassegna di alcuni strumenti di analisi utili alla presentazione dei capitoli che seguiranno. Si rimanda, per approfondimenti, a Angrist e Pischke (2008) e Pearl (2009).

1.1 Il contesto

1.1.1 Il problema fondamentale dell'inferenza causale

Sia Y una variabile d'interesse e si supponga di voler studiare l'effetto che ha su Y un trattamento descritto da una variabile D dicotomica, che assume valore 1 se un soggetto viene trattato e 0 altrimenti. Per caratterizzare il problema in un contesto di inferenza causale si immagini di poter osservare il valore della variabile d'interesse per il soggetto i-esimo, Y_i, sia nel caso in cui questo venga trattato ($D_i = 1$), sia nel caso in cui questo non venga trattato ($D_i = 0$). Si considerano dunque, per ogni individuo, due valori della variabile d'interesse, noti in letteratura come *risultati potenziali*:

$$\begin{aligned} Y_i &= \begin{cases} Y_i^1 & \text{se} \quad D_i = 1 \\ Y_i^0 & \text{se} \quad D_i = 0, \end{cases} \\ &= \quad Y_i^0 + (Y_i^1 - Y_i^0) D_i. \end{aligned} \tag{1.1}$$

In altre parole Y_i^1 è il valore della variabile d'interesse per il soggetto i-esimo se questo si sottopone al trattamento, a prescindere dal fatto che nella realtà

sia trattato o meno, mentre Y_i^0 è il valore per il soggetto i-esimo se non si sottopone al trattamento.

La differenza tra i risultati potenziali di un soggetto, $Y_i^1 - Y_i^0$, rappresenta l'effetto causale del trattamento, che può essere omogeneo nella popolazione o diverso per diversi soggetti. Un tipico indice di sintesi utilizzato per la valutazione del trattamento sulla popolazione è il suo valore atteso, $\mathbb{E}(Y_i^1 - Y_i^0)$ (*average treatment effect*, *ATE*).

Il problema fondamentale dell'inferenza causale è che non è possibile osservare su uno stesso soggetto entrambi i suoi risultati potenziali, ed è quindi necessario ricorrere ad una valutazione dell'effetto del trattamento basato sul confronto tra i soggetti trattati e non trattati che vengono osservati nella popolazione.

1.1.2 Esperimenti randomizzati

Al fine di determinare l'effetto di un trattamento sulla popolazione appare naturale valutare la differenza osservata tra le medie dei valori della variabile d'interesse per i soggetti trattati e non trattati,

$$\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0).$$

Questa quantità, tuttavia, costituisce una rappresentazione attendibile dell'*ATE* solo se è ragionevole assumere che i due gruppi di soggetti trattati e non trattati siano mediamente equivalenti in ogni loro caratteristica e quindi idonei a mimare la differenza media dei risultati potenziali entro i soggetti. Il contesto più credibile per trarre conclusioni di questo tipo è quello degli esperimenti randomizzati (*randomized control trials*, *RCT*). I *RCT* sono studi in cui si dispone di un disegno sperimentale che permette di controllare il processo di selezione dei soggetti al trattamento; in particolare, questa selezione viene fatta in modo casuale.

Da un punto di vista formale, l'assegnazione casuale del valore di D_i per il soggetto i-esimo rende D_i indipendente dai risultati potenziali di Y_i: $D_i \perp \{Y_i^1, Y_i^0\}$. Sotto tale condizione, si ha allora che $\mathbb{E}(Y_i^1|D) = \mathbb{E}(Y_i^1)$, e $\mathbb{E}(Y_i^0|D) = \mathbb{E}(Y_i^0)$ e si può quindi scrivere

$$\begin{aligned}
\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) &= \mathbb{E}(Y_i^1|D_i = 1) - \mathbb{E}(Y_i^0|D_i = 0) \\
&= \underbrace{\mathbb{E}(Y_i^1|D_i = 1) - \mathbb{E}(Y_i^0|D_i = 1)}_{\textit{average treatment effect on the treated}} \\
&= \mathbb{E}(Y_i^1 - Y_i^0|D_i = 1) \\
&= \mathbb{E}(Y_i^1 - Y_i^0),
\end{aligned}$$

ossia la differenza media osservata è pari all'effetto medio del trattamento sui trattati (*average treatment effect on the treated*, *ATT*) che è a sua volta uguale

all'*ATE*. Per questa ragione, quando possibile, i metodi dell'inferenza causale si costruiscono e si valutano considerando i risultati che si otterrebbero in un *RCT* come *benchmark*.

1.1.3 Il *selection bias*

Negli studi osservazionali, che rappresentano, soprattutto negli ambiti socio-economici, la principale metodologia di raccolta dei dati, non c'è controllo sull'assegnazione dei soggetti al trattamento ed è quindi poco credibile che i trattati e non trattati osservati differiscano tra loro solo per il fatto che siano sottoposti o meno al trattamento. È plausibile invece supporre che vi siano altre caratteristiche che differenziano trattati e non trattati e che queste influenzino sia la propensione dei soggetti a sottoporsi al trattamento sia l'entità della loro risposta al trattamento. Queste caratteristiche incidono nel calcolo della differenza tra medie osservate, che quindi non isola il solo effetto causale d'interesse. A livello algebrico si può, infatti, ottenere la seguente scomposizione:

$$\begin{aligned}\mathbb{E}(Y_i|D_i=1)-\mathbb{E}(Y_i|D_i=0) \;&=\; \mathbb{E}(Y_i^1|D_i=1)-\mathbb{E}(Y_i^0|D_i=0) \\ &= \underbrace{\mathbb{E}(Y_i^1|D_i=1)-\mathbb{E}(Y_i^0|D_i=1)}_{ATT} + \underbrace{\mathbb{E}(Y_i^0|D_i=1)-\mathbb{E}(Y_i^0|D_i=0)}_{selection\ bias}.\end{aligned} \tag{1.2}$$

In altre parole, la differenza tra medie osservate può essere scomposta additivamente nell'effetto medio del trattamento sui trattati e un'altra quantità che rappresenta la distorsione da selezione (*selection bias*) ed è data dalla differenza nei risultati potenziali medi di assenza di trattamento tra i gruppi di soggetti trattati e non trattati. Quest'ultima è dovuta alle differenze tra i due gruppi che prescindono dal trattamento oggetto di studio.

Uno dei principali obiettivi dell'inferenza causale è proprio quello di disporre di metodi che eliminino o controllino il *selection bias* in modo da poter identificare correttamente l'effetto del trattamento d'interesse.

1.2 Regressione e causalità

1.2.1 Il modello lineare per valutare un effetto causale

Tra i diversi approcci presenti in letteratura per rispondere a quesiti di natura causale, i modelli di regressione assumono un ruolo cardine, sia in un contesto di *RCT* sia negli studi osservazionali. Le proprietà statistiche di cui gode, ad esempio, il modello lineare, largamente diffuso nella letteratura econometrica, lo rendono infatti uno strumento di analisi attraente per studiare relazioni tra variabili non solo di natura associazionale ma anche di causa-effetto. Tuttavia, perché sia possibile trarre considerazioni di natura causale a partire da un

modello lineare, sono necessarie delle condizioni e delle assunzioni che esulano dalla bontà delle sue proprietà statistiche.

Di seguito, per continuità con la sezione precedente, viene trattato il problema con variabile che descrive il trattamento D dicotomica, ma le considerazioni restano invariate quando la variabile di trattamento assume altra natura, ad esempio politomica o quantitativa.

Assumendo un effetto del trattamento omogeneo, si può riscrivere l'equazione (1.1) in una forma di modello di regressione nel seguente modo:

$$\begin{aligned} Y_i = &\quad Y_i^0 + (Y_i^1 - Y_i^0)D_i \\ = &\quad \underbrace{\mathbb{E}(Y_i^0)}_{\beta_0} + \underbrace{(Y_i^1 - Y_i^0)}_{\beta_1} D_i + \underbrace{Y_i^0 - \mathbb{E}(Y_i^0)}_{\varepsilon_i} \\ = &\quad \beta_0 + \beta_1 D_i + \varepsilon_i, \end{aligned} \tag{1.3}$$

dove β_1 è l'effetto causale d'interesse e il termine di errore ε_i racchiude la variabilità individuale del risultato potenziale in assenza del trattamento, Y_i^0. Pertanto, ciò che si osserva in media per soggetti trattati e non trattati è :

$$\begin{aligned} &\mathbb{E}(Y_i|D_i = 1) = \beta_0 + \beta_1 + \mathbb{E}(\varepsilon_i|D_i = 1) \quad \text{e} \\ &\mathbb{E}(Y_i|D_i = 0) = \beta_0 + \mathbb{E}(\varepsilon_i|D_i = 0), \end{aligned}$$

la cui differenza si può scrivere come

$$\begin{aligned} \mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) \; &= \beta_1 \; + \underbrace{\mathbb{E}(\varepsilon_i|D_i = 1) - \mathbb{E}(\varepsilon_i|D_i = 0)}_{selection\ bias}. \\ &= \beta_1 \; + \mathbb{E}(Y_i^0|D_i = 1) - \mathbb{E}(Y_i^0|D_i = 0), \end{aligned} \tag{1.4}$$

È evidente dall'equazione (1.4) come il *selection bias* dipenda allora dalla correlazione tra la variabile trattamento D_i e il termine di errore ε_i. Si nota inoltre che, per costruzione, parlare di dipendenza tra D_i e ε_i è equivalente a parlare di dipendenza tra D_i e risultati potenziali.

In un *RCT*, dove il trattamento è assegnato casualmente, si ha che $D_i \perp \varepsilon_i$. Pertanto, dalla (1.4) consegue che il *selection bias* si annulla e la differenza media della risposta, tra il gruppo dei trattati e quello dei non trattati, corrisponde all'effetto causale del trattamento, pari al coefficiente β_1, e facilmente stimabile attraverso il metodo dei minimi quadrati. Nel caso di uno studio osservazionale, invece, il trattamento D_i non è assegnato casualmente e il termine di errore racchiude tutta l'*eterogeneità non osservata* tra i soggetti, ovvero assorbe tutte le caratteristiche dei soggetti che differenziano, per propensione e risposta al trattamento, il gruppo dei trattati dai non trattati. Se il valore di D_i dipende dalle caratteristiche del soggetto, chiaramente ε_i è correlata con D_i e il *selection bias* è diverso da zero.

1.2.2 Ipotesi di indipendenza condizionale

Un modo alternativo per definire, in un contesto di regressione, il problema ora emerso, è quello di descrivere l'eterogeneità presente nei soggetti attraverso una o più variabili esplicative.

Si assuma, per semplicità, che tutte e sole le caratteristiche che differenziano il gruppo dei trattati dai non trattati in relazione alla risposta siano definite da una singola variabile Z, che assume valore Z_i nell'$i-$ esimo soggetto, ed è in relazione sia con Y che con D.

Il termine di errore dell'equazione (1.3) può essere scomposto, allora, nella seguente

$$\varepsilon_i = \gamma Z_i + v_i, \tag{1.5}$$

dove $\gamma \in R^p$ è il vettore di coefficienti associato a Z e v_i è un termine casuale a media nulla e incorrelato con Z. Sostituendo l'espressione (1.5) nell'equazione (1.3) il modello diventa:

$$Y_i = \beta_0 + \beta_1 D_i + \gamma Z_i + v_i. \tag{1.6}$$

Pertanto, condizionatamente ad un fissato valore di Z_i si ha che

$$\begin{aligned}
&\mathbb{E}(Y_i|D_i = 1, Z_i) = \beta_0 + \beta_1 + \gamma Z_i + \mathbb{E}(v_i|D_i = 1, Z_i) \quad \text{e} \\
&\mathbb{E}(Y_i|D_i = 0, Z_i) = \beta_0 + \gamma Z_i + \mathbb{E}(v_i|D_i = 0, Z_i).
\end{aligned}$$

Poiché Z_i, per definizione, racchiude tutte e sole le caratteristiche del soggetto che determinano la propensione al trattamento, ne consegue che il termine casuale v_i è incorrelato anche con D_i, ovvero $D_i \perp \{Y_i^1, Y_i^0\}|Z_i$. Quest'ultima rappresenta la formalizzazione di quella che è nota come ipotesi di indipendenza condizionale (*Conditional Independence Assumption, CIA*) e consente, in un contesto di regressione, di interpretare il coefficiente di regressione associato al trattamento in termini di effetto medio causale. Si ha, infatti, che:

$$\begin{aligned}
\mathbb{E}(Y_i|Z_i, D_i = 1) - \mathbb{E}(Y_i|Z_i, D_i = 0) &= \mathbb{E}(Y_i^1|Z_i, D_i = 1) - \mathbb{E}(Y_i^0|Z_i, D_i = 0) \\
&= \mathbb{E}(Y_i^1 - Y_i^0|Z_i) = \beta_1,
\end{aligned}$$

ovvero il *selection bias* sparisce e la differenza in media tra la risposta nei trattati e non trattati definisce l'effetto causale, stimabile in maniera non distorta attraverso il metodo dei minimi quadrati (OLS).

Se, al contrario, non si tiene conto delle caratteristiche Z e si determina la stima OLS del parametro β_1 nel modello (1.3), con opportuni calcoli è agevole mostrare che (Wilms *et al.*, 2021)

$$\mathbb{E}(\hat{\beta}_1) = \beta_1 + \gamma \delta_{Z,D}$$

ovvero la stima dell'effetto del trattamento è distorta, con una distorsione (*omitted variable bias*, OVB) che dipende da γ e da una quantità $\delta_{Z,D}$ che

dipende dalla correlazione tra il trattamento e la variabile omessa. Per approfondimenti, si veda Wilms *et al.* (2021).

Alla luce delle considerazioni emerse appare chiaro che, una strategia per interpretare il coefficiente di regressione associato al trattamento in termini di effetto medio causale, ottenendone una stima non distorta è quella di identificare tutte quelle caratteristiche osservabili, ritenute responsabili della propensione e la risposta al trattamento. In questo modo, condizionandosi a tali caratteristiche, è possibile un confronto tra soggetti simili trattati e non trattati. Tali caratteristiche vengono incluse nel modello con il solo scopo di spiegare una parte di variabilità del fenomeno ausiliaria a quella di specifico interesse, senza un diretto interesse circa il loro effetto sulla variabile risposta, e per questo vengono dette *variabili di controllo*. Per un maggiore approfondimento si veda Angrist e Pischke (2008).

Nel seguito si indicherà con $X_i = (X_{i1}, \ldots, X_{ip})^T$ il vettore di variabili di controllo nel soggetto i.

1.2.3 Complementi

Sebbene l'uso di variabili di controllo consenta, grazie alla CIA, di interpretare il coefficiente di regressione in termini di effetto causale, non sempre è possibile identificare, ed osservare, tutti i fattori responsabili delle differenze tra il gruppo di trattati e non trattati.

In tal caso si incorre nel problema menzionato di variabili omesse, il *selection bias* non si annulla e le stime dell'effetto causale sono distorte e fuorvianti.

Nella letteratura econometrica è stata sviluppata una pluralità di metodologie che in aggiunta, o in alternativa, all'inserimento di variabili di controllo, mirano a ridurre la distorsione. Fare una rassegna completa di tali approcci esula dagli scopi di questa tesi e si rimanda per una trattazione completa, a Angrist e Pischke (2008).

Restando nel contesto dei modelli di regressione, si menzionano, tra gli altri, il metodo della *Regression discontinuity*, utilizzata quando il trattamento in analisi consiste in una regola, nota, legata ad una soglia di una determinata caratteristica che suddivide la popolazione tra trattati e non trattati in modo più o meno netto, o i modelli a effetti fissi, *first differences*, *differences-in-differences*, che si utilizzano per lo più quando si hanno a disposizione dati di tipo panel e si studia l'effetto del trattamento sulle differenze di *outcome* entro i soggetti stessi.

Una metodologia che merita una particolare attenzione nell'inferenza causale è l'uso delle cosiddette variabili strumentali, che risultano particolarmente utili nei casi in cui non si hanno sufficienti controlli da poter considerare credibile la *CIA*. Si definisce variabile strumentale una variabile correlata alla variabile di trattamento ma incorrelata con qualsiasi altra variabile che abbia un'influenza

sulla risposta. In altre parole, una variabile strumentale è indipendente dai risultati potenziali ed ha un effetto sull'*outcome* solo attraverso il trattamento.

Concettualmente, la stima dell'effetto di un trattamento con inclusione di una variabile strumentale si compone solitamente di due passi, con metodi noti come stimatore di Wald o *two-stage least squares* (2SLS). Nel caso del 2SLS, ad esempio, il primo passo (*first stage*), è un modello che mette in relazione la variabile di trattamento con lo strumento, ed è volta a quantificare la correlazione tra le due variabili; il secondo (*second stage*), è una regressione come (1.6) in cui, come esplicativa, al posto dei valori osservati della variabile trattamento si inseriscono i suoi valori previsti al primo stadio; intuitivamente, in questo modo si misura l'effetto sull'outcome della sola parte di variabilità del trattamento dovuta da una variazione quasi-sperimentale, ossia dovuta allo strumento, ed è perciò una stima non affetta da *selection bias*.

Va tuttavia considerato che le assunzioni che permettono di considerare una variabile uno strumento sono piuttosto restrittive e rendono questa metodologia spesso non attuabile. Per approfondimenti si veda Imbens (2014).

1.3 Effetti eterogenei

Finora è stato affrontato il problema della valutazione dell'effetto di un trattamento assumendo che questo sia omogeneo nella popolazione. In alcuni contesti, tuttavia, è plausibile presumere che ci sia un'eterogeneità tra soggetti anche nella loro risposta al trattamento, ad esempio per diversi sottogruppi di popolazione.

Cogliere questo tipo di eterogeneità è di fondamentale importanza. Si pensi, ad esempio, alla situazione in cui siano presenti nella popolazione gruppi di soggetti che rispondono ad un trattamento anche in maniera opposta tra loro. Ignorare tale eterogeneità può condurre ad una compensazione degli effetti qualora si utilizzi un indice di sintesi per l'intera popolazione.

In questi casi, quindi, più che la valutazione dell'ATE, è allora di interesse la valutazione di una sua versione condizionata a determinate tipologie di soggetti. Si definisce quindi il *conditional average treatment effect* ($CATE$) come la differenza media tra i risultati potenziali di una variabile d'interesse limitatamente ad alcuni soggetti con determinate caratteristiche. A scopo illustrativo, si consideri ancora una volta una situazione semplificata in cui tali caratteristiche siano descritte da una variabile Z binaria. Formalmente si avrà:

$$CATE(Z) = \mathbb{E}(Y_i^1 - Y_i^0 | Z_i).$$

Una valutazione del $CATE$ può essere ottenuta riconducendo il problema, ancora una volta, ad uno schema di regressione. Sia inoltre X un vettore di variabili di controllo, per il quale sia soddisfatta la CIA. È possibile allora af-

frontare il problema mediante stratificazione dei soggetti sulla base dei valori di Z:

$$\begin{cases} Y_i = \beta_{00} + \beta_{10} D_i + \varepsilon_i & \text{se} \quad Z_i = 0 \\ Y_i = \beta_{01} + \beta_{11} D_i + \varepsilon_i & \text{se} \quad Z_i = 1 \end{cases} \tag{1.7}$$

dove β_{1z} rappresenta il CATE, $z \in \{0, 1\}$.

Similmente, la valutazione del CATE può coinvolgere uno schema di regressione in cui Z interagisce con il trattamento determinandone effetti diversi. Formalmente:

$$Y_i = \beta_0 + \beta_1 D_i + \alpha_0 Z_i + \alpha_1 D_i Z_i + \varepsilon_i \tag{1.8}$$

dove, ponendo $\beta_0 = \beta_{00}$, $\beta_1 = \beta_{10}$, $\beta_0 + \alpha_0 = \beta_{01}$ e $\beta_1 + \alpha_1 = \beta_{11}$ si riottiene una formulazione equivalente al modello (1.7).

Negli anni recenti, la studio di metodologie per la valutazione del *CATE* ha ricevuto particolare enfasi, soprattutto con riferimento all'opportunità di utilizzare alcuni strumenti di *machine learning*. Si vedano, a titolo di esempio, Duflo (2018); Chernozhukov *et al.* (2023).

Capitolo 2

Modelli di regressione a mistura finita

I modelli mistura sono una classe di modelli statistici molto flessibili che permettono di rappresentare e analizzare dati complessi. L'utilizzo di questi modelli è motivato in contesti in cui si assume che i dati osservati siano realizzazioni di un processo generatore caratterizzato da una fonte non osservabile di eterogeneità.

In questo capitolo si discuteranno i principali aspetti che riguardano la formulazione, l'interpretazione e l'inferenza su un modello a mistura finita, concentrando l'attenzione in particolare sui modelli di regressione. La trattazione fa riferimento principalmente a McLachlan e Peel (2004, Cap. 5) e Bouveyron *et al.* (2019, Cap. 11), cui si rimanda per approfondimenti.

2.1 Specificazione del modello

2.1.1 Modellazione dell'eterogeneità mediante variabili latenti

In molti contesti reali, e soprattutto con riferimento alle applicazioni socio-economiche, è ragionevole assumere che il fenomeno di interesse si manifesti in modo eterogeneo all'interno della popolazione, ad esempio perché riconducibile alla presenza di sottopopolazioni definite a partire da una o più caratteristiche non osservate, o non osservabili, e per questo dette latenti.

Formalmente, sia $Z = (Z_1, \ldots, Z_G)$ un vettore di variabili binarie, con $Z_g = 1$ nella sottopopolazione latente $g, g = 1, \ldots, G$, 0 altrimenti. L'eterogeneità della variabile di interesse Y può essere descritta assumendo che

$$Y|Z_g = 1 \sim p_g(y; \theta_g), \qquad g = 1, \ldots, G$$

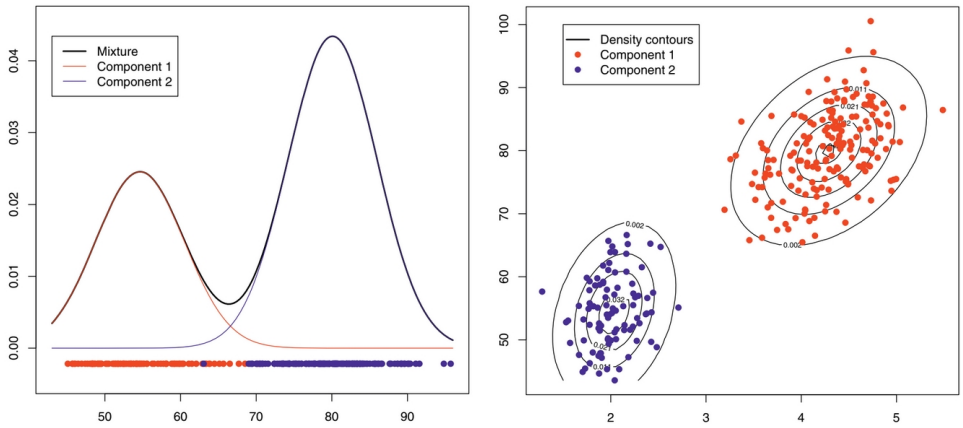


Figura 2.1: Esempio di mistura finita a due componenti gaussiane in $\mathbb{R}$ (sinistra) e $\mathbb{R}^2$ (destra). Le componenti della mistura, moltiplicate per il loro peso, sono indicate in rosso e blu, mentre la densità risultante dalla loro somma è rappresentata in nero. Cfr. Figure 2.1 e 2.2 in Bouveyron *et al.* (2019).

dove, a seconda della natura di Y, p_g è una funzione di probabilità o densità, tipicamente appartenente ad una famiglia parametrica governata da un vettore θ_g di parametri. In altre parole, Y è descritta da un modello distributivo diverso a seconda della sottopopolazione di appartenenza.

Marginalizzando rispetto a Z, la distribuzione della variabile di interesse diventa:

$$\begin{aligned} p(y) &= \sum_{g=1}^{G} p(y, z) = \sum_{g=1}^{G} p(Z_g = 1) p(y|\theta_g, Z_g = 1) \\ &= \sum_{g=1}^{G} \pi_g p_g(y|\theta_g), \end{aligned} \tag{2.1}$$

che rappresenta l'equazione di un modello a mistura finita a G componenti. I pesi (o *mixing proportions*) $\pi_g = \mathbb{P}(Z_g = 1)$ sono le probabilità (a priori) che un'osservazione sia generata dalla g-esima sottopopolazione e sono quindi tali che $\pi_g \geq 0$, $g = 1, \ldots, G$, e $\sum_{g=1}^{G} \pi_g = 1$. Le distribuzioni $p_g(y_i|\theta_g)$, $g = 1, \ldots, G$, sono dette componenti della mistura (*mixture components*). Si veda la Figura 2.1 per un'illustrazione con $G = 2$ e $Y \in \mathbb{R}$ e $\mathbb{R}^2$.

Sebbene il modello (2.1) sia generalizzabile al caso di infinite componenti, è spesso sufficiente lavorare con modelli a mistura finita, anche tenendo conto che ogni mistura infinita di distribuzioni può essere approssimata ragionevolmente bene da un numero finito di componenti (Titterington *et al.*, 1985). Nella trattazione che segue, G sarà assunto noto, e l'ipotesi verrà successivamente rilassata nel paragrafo 2.3.

2.1.2 Misture finite di modelli di regressione

Quando è ragionevole supporre che, oltre alle fonti di eterogeneità non osservabili, siano disponibili anche alcuni fattori concomitanti $X = (X_1, \ldots, X_p)^T$ che esercitano un'influenza sulla variabile risposta, i modelli a mistura finita diventano uno strumento utile e flessibile per lo studio della distribuzione condizionata di $Y|X$. In tal caso, si assume che

$$Y|Z_g = 1, X = x \sim p_g(y|x, \theta_g),\ g = 1, \ldots, G$$

e la (2.1) si estende come segue:

$$\begin{aligned} Y|X = x \sim p(y|x) &= \sum_{g=1}^{G} p(y_i, z_i|x_i) \\ &= \sum_{g=1}^{G} p(Z_g = 1) p(y_i|x_i, \theta_g, Z_g = 1) \\ &= \sum_{g=1}^{G} \pi_g p_g(y|x, \theta_g) \qquad (2.2) \\ &= \begin{cases} p_1(y|x, \theta_1) & \text{con probabilità } \pi_1 \\ \vdots & \\ p_G(y|x, \theta_G) & \text{con probabilità } \pi_G \end{cases} \qquad (2.3) \end{aligned}$$

che rappresenta un modello di regressione a mistura finita. Le componenti della mistura possono descrivere variabili di qualsiasi natura (continua o discreta) ma comunemente la scelta ricade tra le distribuzioni appartenenti alla famiglia esponenziale in modo da poter ricondurre agevolmente lo studio della relazione tra la variabile di interesse e le variabili esplicative alla teoria dei modelli lineari generalizzati. Nel caso più semplice, $p_g(y|x, \theta_g) = \phi(y|x, \mu_g, \sigma_g^2)$ sono distribuzioni normali, e la media, $\mu_g(x) = \mu_g$, è legata linearmente alle variabili esplicative dall'equazione

$$\mu_g = \beta_{0g} + x^T \beta_{1g},$$

dove $\beta_{0g} \in \mathbb{R}$ e $\beta_{1g} \in \mathbb{R}^p$ sono parametri ignoti. Indicata con Y_i una generica copia di Y, e con $Z_i = (Z_{11}, \ldots, Z_{iG})$ il vettore associato di variabili latenti, il modello di riferimento può riscritto nella forma classica di un modello di regressione lineare:

$$Y_i = \begin{cases} \beta_{01} + x_i^T \beta_{11} + \varepsilon_{i1} & \text{se } Z_{i1} = 1 \text{ (con probabilità } \pi_1) \\ \vdots & \\ \beta_{0G} + x_i^T \beta_{1G} + \varepsilon_{iG} & \text{se } Z_{iG} = 1 \text{ (con probabilità } \pi_G), \end{cases} \qquad (2.4)$$

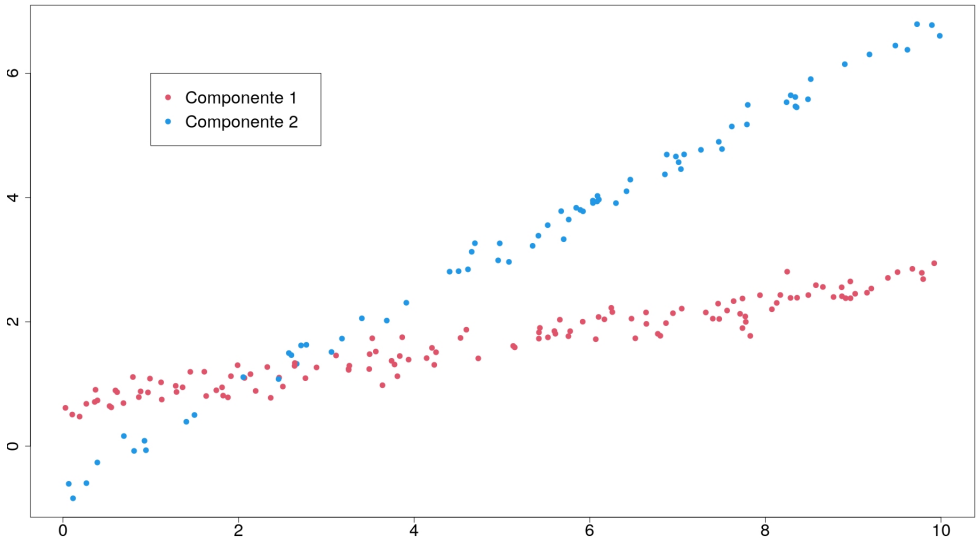


Figura 2.2: Esempio di mistura di modelli di regressione a due componenti (rispettivamente in rosso e in blu).

con $\varepsilon_{ig} \sim N(0, \sigma_g^2)$, $g = 1, \ldots, G$.

È facile allora vedere come, in questa specificazione, un modello mistura di regressione di Y rispetto a X rappresenti, di fatto, un modello di regressione con parametri diversi per ciascuna componente della mistura. Esso può essere dunque interpretato come modello a componenti eteroschedastiche che include le componenti di Z quali ulteriori variabili esplicative, ancorché non osservate, ed eventualmente interagenti con le X. Ad esempio, se $G = 2$:

$$Y_i = \beta_{01} + x_i^T \beta_{11} + \alpha_0 Z_{i2} + \alpha_1 x_i^T Z_{i2} + \varepsilon_i,$$

dove $\alpha_0 \in \mathbb{R}$ e $\alpha_1 \in \mathbb{R}^p$, è una specificazione equivalente al modello (2.4) con $G = 2$ dove $\beta_{02} = \beta_{01} + \alpha_0$ e $\beta_{12} = \beta_{11} + \alpha_1$. La Figura 2.2 rappresenta un'illustrazione di un modello a due componenti.

È possibile estendere ulteriormente il modello (2.2) mettendo in relazione i fattori concomitanti X con la variabile latente Z che determina le componenti della mistura. In questo modo, il potere esplicativo delle concomitanti non si limita al comportamento della variabile d'interesse ma ha impatto anche sulla probabilità che un'osservazione venga generata da una specifica componente. Dal punto di vista modellistico questo significa specificare per la variabile latente Z una distribuzione condizionata alle X. Supponendo che Z possa essere ben rappresentata da una variabile con distribuzione multinomiale, si può specificare il modello come segue:

$$\begin{cases} Y|Z_g = 1, X = x \sim p_g(y|x, \theta_g) \\ Z|X = x \sim Bi_G(1, \pi) \end{cases} \tag{2.5}$$

dove il vettore di probabilità $\pi = \pi(x)$, viene messo in relazione alle variabili esplicative come in un modello di regressione multinomiale, attraverso il predittore lineare:

$$\log\frac{\pi_g}{\pi_1} = \gamma_{0g} + x^T\gamma_{1g} \quad g = 2, \ldots, G,$$

con $\gamma_{0g} \in R$ e $\gamma_{1g} \in R^p$ parametri ignoti.
Marginalizzando la distribuzione di Y rispetto a Z si ottiene

$$\begin{aligned} p(y|x) &= \sum_{g=1}^{G} p(y, z|x) \\ &= \sum_{g=1}^{G} p(Z_g = 1|x)\, p(y|x, Z_g = 1) \\ &= \sum_{g=1}^{G} \pi_g(x)\, p_g(y|x, \theta_g), \end{aligned} \tag{2.6}$$

conosciuto in letteratura come modello a mistura degli esperti (*mixture-of-experts model*). È possibile specificarne diverse versioni a seconda dell'esclusione o inclusione delle variabili esplicative nelle due distinte parti che compongono la mistura (pesi e componenti). Per una rassegna completa si veda Gormley e Murphy (2011).

2.2 Inferenza

2.2.1 Stima di Massima Verosimiglianza

Sebbene, in letteratura, siano discussi diversi metodi per la stima di modelli mistura, tra i quali riveste un ruolo sempre crescente l'approccio Bayesiano (si veda, ad esempio, Fruhwirth-Schnatter *et al.*, 2019), la massima verosimiglianza rappresenta, almeno storicamente, lo strumento principale per l'inferenza. Nel seguito si presenta una sintesi dei principali aspetti relativi all'utilizzo di tale approccio, facendo in particolare riferimento ai modelli di regressione.

Siano $(y_1, x_1), \ldots, (y_n, x_n)$ i dati osservati, $(y_i, x_i) \in R^{p+1}$ per $i = 1, \ldots, n$, e si assuma che questi siano realizzazioni indipendenti da una distribuzione mistura come quella riportata in equazione (2.2). Si può quindi definire, a partire dai

dati osservati, una funzione di verosimiglianza *osservata* per i parametri del modello $(\pi, \theta) = (\pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G)$ del tipo:

$$L^O(\pi, \theta | y, x) = \prod_{i=1}^{n} p(y_i | x_i) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g p_g(y_i | x_i, \theta_g), \tag{2.7}$$

con rispettiva log-verosimiglianza osservata definita come

$$\ell^O(\pi, \theta | y, x) = log\left[L^O(\pi, \theta | y, x)\right] = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g p_g(y_i | x_i, \theta_g).$$

La presenza del logaritmo di una somma rende impossibile risolvere analiticamente le equazioni di verosimiglianza, a prescindere dai modelli distributivi assunti. Un algoritmo utile a risolvere elegantemente il problema dell'ottimizzazione numerica è noto con il nome di *Expectation-Maximization* (EM, Dempster *et al.* (1977)) e coinvolge la definizione di una seconda funzione di verosimiglianza, detta *completa* perché formulata immaginando di aver osservato i dati *completi* $(y_1, x_1, z_1), \ldots, (y_n, x_n, z_n)$, dove le z_i, per $i = 1, \ldots, n$, sono realizzazioni *iid* della variabile $Z_i = (Z_{i1}, \ldots, Z_{iG}) \sim Bi_G(1, \pi = (\pi_1, \ldots, \pi_G))$:

$$L^C(\pi, \theta | y, x, z) = \prod_{i=1}^{n} p(y_i, z_i | x_i) = \prod_{i=1}^{n} p(z_i) p(y_i | z_i, x_i) \tag{2.8}$$

$$= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g^{z_{ig}}\right] \prod_{g=1}^{G} \left[p_g(y_i | x_i, \theta_g)\right]^{z_{ig}} = \tag{2.9}$$

$$= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g p_g(y_i | x_i, \theta_g)\right]^{z_{ig}}, \tag{2.10}$$

con rispettiva log-verosimiglianza *completa*:

$$\ell^C(\pi, \theta | y, x, z) = \sum_{i=1}^{n} \sum_{g=1}^{G} \log \left[\pi_g p_g(y_i | x_i, \theta_g)\right]^{z_{ig}} \tag{2.11}$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \left[\pi_g p_g(y_i | x_i, \theta_g)\right]. \tag{2.12}$$

2.2.2 L'Algoritmo EM

L'algoritmo EM è stato originariamente formulato per determinare una stima di massima verosimiglianza in presenza di dati mancanti. Data l'inosservabilità

della classe latente il suo utilizzo è particolarmente appropriato nel caso di modelli mistura dove le Z_i vengono trattate, appunto, alla stregua di dati mancanti.

L'idea alla base dell'algoritmo è quella di ottimizzare, anziché la (2.12), il suo valore atteso al variare di Z.

$$\begin{aligned} Q(\pi, \theta; y, x) &= \mathbb{E}_Z \left[\ell^C(\pi, \theta | y, x) \right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}_Z \left[Z_{ig} | y, x \right] \log \left[\pi_g p_g(y_i | x_i, \theta_g) \right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig} \log \left[\pi_g p_g(y_i | x_i, \theta_g) \right] \end{aligned} \tag{2.13}$$

dove, per $g = 1, \ldots, G$ e per $i = 1, \ldots, n$,

$$\begin{aligned} \tau_{ig} = \mathbb{E}_Z \left[Z_{ig} | y, x \right] &= P(Z_{ig} = 1 | y, x) \\ &= \frac{p(Z_{ig} = 1) p_g(y_i | x_i, \theta_g, Z_{ig} = 1)}{\sum_{g=1}^{G} p(Z_{ig} = 1) p_g(y_i | x_i, \theta_g, Z_{ig} = 1)} \\ &= \frac{\pi_g p_g(y_i | x_i, \theta_g)}{\sum_{g=1}^{G} \pi_g p_g(y_i | x_i, \theta_g)}. \end{aligned}$$

La quantità τ_{ig} corrisponde alla probabilità a posteriori che l'i-esimo soggetto appartenga alla g-esima componente.

Supposti noti i valori di τ_{ig} la massimizzazione della funzione $Q(\cdot)$ diventa particolarmente agevole, perché corrisponde alla massimizzazione dei contributi individuali a ciascuna delle componenti della mistura, pesati rispetto alle probabilità a posteriori di appartenervi. Per questa ragione l'algoritmo affronta il problema di ottimizzazione iterativamente, alternando due passi noti appunto come *Expectation* (E) e *Maximization* (M).

Posto $(\pi^{(k)}, \theta^{(k)})$ il vettore dei parametri all'iterazione (k), all'iterazione $(k+1)$ l'E-step determina la (2.13) per quella fissata configurazione di parametri:

$$\tau_{ig}^{(k+1)} = \frac{\pi_g^{(k)} p_g(y_i | x_i, \theta_g^{(k)})}{\sum_{g=1}^{G} \pi_g^{(k)} p_g(y_i | x_i, \theta_g^{(k)})}.$$

L'M-step è il passo di aggiornamento delle stime dei parametri ignoti e consiste nella massimizzazione della funzione $Q(\cdot)$ rispetto a (π, θ), tenendo fissate le probabilità a posteriori $\tau_{ig}^{(k+1)}$ calcolate nell'E-step:

$$
\begin{aligned}
(\pi^{(k+1)}, \theta^{(k+1)}) &= \underset{(\pi,\theta)}{\operatorname{argmax}}\ Q(\pi, \theta; \pi^{(k)}, \theta^{(k)}) \\
&= \underset{(\pi,\theta)}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log\ [\pi_g p_g(y_i|x_i, \theta_g)] \\
&= \underset{(\pi,\theta)}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log p_g(y_i|x_i, \theta_g)
\end{aligned}
\tag{2.14}
$$

Si noti dall'espressione (2.14) che la funzione $Q(\cdot)$ può essere scomposta additivamente in due componenti che permettono di trovare separamente i punti di massimo per i parametri π e θ. Si ha quindi che, per i pesi della mistura:

$$
\pi^{(k+1)} = \underset{\pi}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log \pi_g \ \Leftrightarrow\ \pi_g^{(k+1)} = \sum_{i=1}^{n} \tau_{ig}^{(k+1)}/n, \quad g = 1, \ldots, G
$$

l'aggiornamento del parametro di ogni componente consiste in una media delle probabilità a posteriori per quella componente calcolate all'E-step del passo $(k+1)$-esimo. Per quanto riguarda i parametri delle componenti della mistura si ha

$$
\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log p_g(y_i|x_i, \theta_g).
$$

Un aspetto vantaggioso di questa procedura è che solitamente anche per l'aggiornamento del parametro θ si hanno soluzioni in forma chiusa. Nel caso in cui, ad esempio, si assumessero delle componenti della mistura Normali, l'aggiornamento dei parametri si ottiene mediante minimi quadrati pesati (WLS) per le probabilità a posteriori $\tau_{ig}^{(k+1)}$.

Alternando ripetutamente E-step ed M-step la verosimiglianza osservata aumenta monotonicamente e si arriva a convergenza dell'algoritmo quando l'incremento di verosimiglianza tra un'iterazione e quella successiva è sotto una soglia ε prestabilita:

$$
(\hat{\pi}, \hat{\theta}) = (\pi^{(k+1)}, \theta^{(k+1)}) \ \Leftrightarrow\ \left\{ L^O(\pi^{(k+1)}, \theta^{(k+1)}) - L^O(\pi^{(k)}, \theta^{(k)}) \right\} < \varepsilon.
$$

I valori dei parametri a cui l'algoritmo converge costituiscono un punto di massimo locale della verosimiglianza osservata definita per un modello mistura. Questa, tuttavia, non è una funzione globalmente concava e presenta

tipicamente un andamento irregolare con diversi punti di massimo locale. Inoltre sulla frontiera dello spazio parametrico esplode ad infinito. Questi motivi rendono di cruciale importanza la scelta dei valori dei parametri con cui inizializzare l'algoritmo. Esistono diversi approcci per individuare i valori iniziali dei parametri e possono prevedere partizioni casuali della popolazione osservata o l'utilizzo di metodi di raggruppamento parametrici o non parametrici (si veda Bouveyron *et al.* (2019)).

2.2.3 Valutazione dell'incertezza delle stime

Una tipica criticità dell'algoritmo EM è che, per sua costruzione, non fornisce una diretta stima della matrice di covarianza delle stime di massima verosimiglianza. Sono stati proposti diversi metodi per ottenere un calcolo della variabilità delle stime che si ottengono dall'applicazione dell'algoritmo EM, la maggior parte dei quali si basano o sulla matrice d'informazione osservata o su metodi di ricampionamento. Per una rassegna, si veda McLachlan e Peel (2004).

I metodi basati sulla matrice d'informazione osservata derivano direttamente dal risultato teorico di efficienza asintotica degli stimatori di massima verosimiglianza che, sotto condizioni di regolarità, è noto abbiano distribuzione asintoticamente normale con matrice di covarianza pari all'inversa dell'informazione attesa (Azzalini, 2001). È allora possibile avere una valutazione della matrice di covarianza dello stimatore calcolando l'inversa della matrice d'informazione osservata, $I^O(\hat{\pi}, \hat{\theta})$, formalmente definita come l'Hessiana della log-verosimiglianza osservata calcolata nel punto di massima verosimiglianza e cambiata di segno. Definendo le quantità di derivata prima (*score*) e derivata seconda osservate come

$$U^O(\pi, \theta) = \frac{\partial \ell^O(\pi, \theta)}{\partial(\pi, \theta)} \quad \text{e} \quad H^O(\pi, \theta) = \frac{\partial U^O(\pi, \theta)}{\partial(\pi, \theta)},$$

si calcola la matrice d'informazione osservata come $I^O(\hat{\pi}, \hat{\theta}) = -H^O(\hat{\pi}, \hat{\theta})$. Si possono definire allo stesso modo le stesse quantità considerando i dati completi, quindi nel caso in cui si consideri la log-verosimiglianza completa $\ell^C(\pi, \theta)$ e si hanno dunque la quantità *score* completa, $U^C(\pi, \theta)$, l'Hessiana completa, $H^C(\pi, \theta)$, e la matrice d'informazione osservata per dati completi, $I^C(\hat{\pi}, \hat{\theta})$.

Nel caso di modelli mistura il calcolo della derivata seconda della log - verosimiglianza osservata può essere particolarmente complesso da un punto di vista algebrico. Si può dimostrare tuttavia (Louis 1982) che l'Hessiana della log-verosimiglianza osservata cambiata di segno, $I^O(\pi, \theta)$, è legata alla sua controparte definita per dati completi dalla seguente espressione:

$$
\begin{aligned}
I^O(\pi,\theta) &= I^C(\pi,\theta) - Cov\left\{U^C(\pi,\theta)|y,x\right\} \\
&= I^C(\pi,\theta) - E\left\{U^C(\pi,\theta)U^C(\pi,\theta)^T|y,x\right\} + U^O(\pi,\theta)U^O(\pi,\theta)^T,
\end{aligned}
$$

dove l'ultimo passaggio segue dal fatto che $E\left\{U^C(\pi,\theta)|y,x\right\} = U^O(\pi,\theta)$. Poiché, per definizione, nel punto di massima verosimiglianza si ha che $U^O(\hat{\pi},\hat{\theta}) = 0$ allora si può scrivere la matrice d'informazione osservata nel seguente modo:

$$
I^O(\hat{\pi},\hat{\theta}) = I^C(\hat{\pi},\hat{\theta}) - E\left\{U^C(\pi,\theta)U^C(\pi,\theta)^T|y,x\right\}|_{(\pi,\theta)=(\hat{\pi},\hat{\theta})}. \tag{2.15}
$$

In questo modo la matrice d'informazione osservata può essere calcolata attraverso quantità derivanti dalla specificazione del problema con dati completi, che consente una scrittura più trattabile delle quantità di verosimiglianza. Il calcolo può essere ulteriormente semplificato nel caso in cui le componenti della mistura abbiano distribuzioni appartenenti alla famiglia esponenziale, poiché si ha che $I^C(\hat{\pi},\hat{\theta})$ non dipende da Y. Una volta calcolata la matrice d'informazione osservata gli errori standard delle stime dei parametri si possono direttamente ricavare dalla diagonale della sua inversa.

Per alcuni modelli mistura più complessi può risultare complicato calcolare la matrice d'informazione osservata anche attraverso l'espressione (2.15). In questi casi si può ricorrere ad altre metodologie come il *Supplemented EM algorithm* proposto da Meng e Rubin (1989) o metodi computazionalmente più sofisticati che si basano su tecniche di *jackknife* o di ricampionamento *bootstrap* (si veda, ad esempio, O'Hagan et al. 2019).

2.3 Selezione del modello

Per un modello mistura di regressione ci sono diversi aspetti della sua specificazione che necessitano di una fase di confronto e di selezione del modello migliore tra più modelli adattati ai dati. Oltre alla selezione delle covariate utili, sono di rilevante importanza gli aspetti di scelta delle distribuzioni delle componenti di mistura e di numero di componenti del modello. Per questi due ultimi aspetti esiste una sorta di *trade-off*: un modello con distribuzioni delle componenti di mistura più semplici e poco flessibili necessita di più componenti per adattarsi bene ai dati mentre, viceversa, per un modello specificato con singole componenti caratterizzate da distribuzioni più flessibili è sufficiente un numero minore di componenti.

Un semplice metodo grafico per valutare la bontà di adattamento ai dati dei modelli stimati può essere il confronto tra la distribuzione delle osservazioni della variabile risposta condizionata ad alcuni valori delle covariate e la densità condizionata del modello adattato.

Un criterio diffuso per la selezione del modello nell'ambito dei modelli mistura segue un approccio di tipo Bayesiano. L'idea alla base è che, se si mettono a confronto K modelli, $M_1, \ldots, M_K$, con probabilità a priori che siano il modello corretto $p(M_k)$, $k = 1, \ldots, K$ (spesso assunta uniforme), allora, attraverso il teorema di Bayes, la probabilità a posteriori che il modello M_k sia corretto, dati i dati osservati (y, x) è proporzionale alla probabilità che si osservino i dati (y, x) dato il modello M_k moltiplicata alla probabilità a priori per il modello M_k. In formule:

$$p(M_k|y, x) \propto p(y|x, M_k)p(M_k).$$

Selezionare il modello migliore con un approccio Bayesiano significa scegliere il modello con maggior probabilità a posteriori. Se sono presenti parametri ignoti, θ, allora la probabilità $p(y|x, M_k)$ può essere calcolata marginalizzando rispetto ad essi:

$$p(y|x, M_k) = \int p(y|x, M_k, \theta_{M_k})p(\theta_{M_k}|x, M_k)d\theta_{M_k}, \quad k = 1, \ldots, K. \tag{2.16}$$

e per questo motivo è detta verosimiglianza marginale del modello M_k. La principale difficoltà dell'utilizzo di questo approccio è il calcolo di (2.16); si dimostra, tuttavia, che per modelli regolari questa ha uno stretto legame con il BIC (*Bayesian Information Criterion*), infatti:

$$-2\log p(y|x, M_k) \approx -2\log p(y|x, M_k, \hat{\theta}_{M_k}) + \nu_{M_k}\log n \; = \mathrm{BIC}_{M_k}, \tag{2.17}$$

dove ν_{M_k} è il numero di di parametri indipendenti stimati nel modello M_k (Schwarz, 1978). Viene selezionato il modello che mostra il valore del BIC più basso tra quelli considerati. I modelli mistura non soddisfano le condizioni di regolarità richieste per la dimostrazione di (2.17), tuttavia, Keribin (1998) ha dimostrato la consistenza del BIC come criterio per la selezione del numero di componenti di un modello mistura (sotto deboli condizioni di regolarità) e, anche a fronte di procedure alternative più mirate, il BIC è nella pratica il criterio di selezione più diffusamente usato ed efficace per la selezione di un modello mistura. Si veda, per approfondimenti, Celeux *et al.* (2018).

Capitolo 3

Modelli Mistura per la stima di un effetto causale

Una volta illustrate le principali criticità legate alla stima di un effetto causale, i limiti degli strumenti tipicamente utilizzati per affrontare il problema, e presentati i vantaggi dell'uso di modelli mistura come metodologia per identificare dai dati una possibile fonte di eterogeneità non osservabile, in questo capitolo si propone una metodologia volta a produrre una sintesi tra i due approcci discussi, che utilizza i modelli mistura per la stima dell'effetto causale.

3.1 Eterogeneità non osservata come problema di dati incompleti

Come evidenziato nel Capitolo 1, l'inclusione di variabili di controllo in un modello di regressione consente di attribuire un'interpretazione causale all'effetto del parametro associato ad un trattamento. Questo è reso possibile da un'ipotesi di indipendenza condizionale (CIA), ovvero assumendo che le variabili di controllo a disposizione permettano di catturare tutta l'eterogeneità che caratterizza e distingue i due gruppi di soggetti, trattati e non trattati, e che quindi, condizionatamente a queste, i risultati potenziali dei soggetti siano indipendenti dall'assegnazione al trattamento.

Se è vero che la sempre crescente disponibilità di dati e variabili rappresenta una rassicurazione in merito alla validità della CIA, è spesso verosimile supporre che, anche condizionatamente alle caratteristiche osservate, esista una fonte residua di eterogenità non osservabile. In tal caso si incorre nel già menzionato problema di variabili omesse, il *selection bias* non si annulla e le stime dell'effetto causale sono distorte e fuorvianti.

D'altra parte, come evidenziato nel Capitolo 2, i modelli mistura rappresentano uno strumento utile proprio nelle situazioni in cui il fenomeno di interesse sia

caratterizzato da una fonte non osservabile di eterogeneità, e si caratterizzano per affrontare il problema secondo una prospettiva per cui tale eterogeneità è espressa proprio da una variabile mancante.

L'obiettivo che ci si pone è allora quello di esplorare la possibilità di identificare, a partire dai dati stessi, l'eterogeneità residua, mediante il ricorso a un modello di regressione a mistura finita, e avere così una migliore caratterizzazione della popolazione in esame ed una conseguente migliore stima dell'effetto del trattamento.

Alla base del metodo, da un punto di vista della natura causale del quesito d'interesse, vi è l'assunzione che, al netto di eventuali altre caratteristiche osservate nei soggetti, siano presenti nella popolazione dei gruppi, di cui è responsabile una caratteristica non osservata, entro i quali il trattamento si può considerare *as good as randomly assigned* e si possa quindi ottenere entro questi gruppi delle stime corrette del suo effetto.

Formalmente, si consideri una variabile di interesse Y, sulla quale si vuole valutare l'effetto di un trattamento D. Per fissare le idee, in quanto segue si assumerà che $Y \in \mathbb{R}$ e che $D \in \{0, 1\}$ con la doverosa precisazione che tale scelta è unicamente motivata da ragioni di chiarezza espositiva e tale ipotesi non è in alcun modo vincolante. Si indichi inoltre con (X, Z) un vettore di caratteristiche ausiliarie, che si ritiene possano cogliere ed esaurire tutta l'eterogeneità della popolazione al variare del trattamento. Coerentemente con l'impianto classico dei metodi di inferenza causale, si assumerà allora che valga un'ipotesi di indipendenza condizionata, sicché, per ogni soggetto i

$$D_i \perp \{Y_i^1, Y_i^0\} | (X_i, Z_i),$$

e quindi, in analogia con quanto discusso nel Capitolo 1,

$$\begin{aligned} \mathbb{E}(Y_i|X_i, Z_i, D_i = 1) - \mathbb{E}(Y_i|X_i, Z_i, D_i = 0) &= \mathbb{E}(Y_i^1|X_i, Z_i) - \mathbb{E}(Y_i^0|X_i, Z_i) \\ &= \mathbb{E}(Y_i^1 - Y_i^0|X_i, Z_i) \end{aligned}$$

Tuttavia, la natura di tali caratteristiche è tale per cui $X = (X_1, \ldots, X_p)^T$ è un vettore di variabili di controllo, note ed osservabili, di qualunque tipo (categoriali o quantitative) mentre Z è una variabile non osservabile, e dunque latente. Poiché Z raccoglie ed esaurisce tutta l'eterogeneità residua, ha senso una sua descrizione categoriale sconnessa, in modo che modalità diverse siano in grado di cogliere fattori di confondimento che si muovono anche in direzioni diverse. Come nel Capitolo 2 si porrà allora $Z = (Z_1, \ldots, Z_G)$ dove $Z_g = 1$ in corrispondenza della modalità, o gruppo $g, g = 1, \ldots, G$, 0 altrimenti.

3.2 Specificazione del modello

La metodologia che si propone prevede la specificazione di un modello di regressione che, condizionatamente al gruppo di appartenenza, descrive la relazione tra Y e il trattamento D, al netto delle variabili di controllo. In quanto segue, la procedura è illustrata con riferimento ad un modello lineare, ma è facilmente estendibile a specificazioni alternative, come verrà discusso in seguito. Si avrà allora

$$Y|(Z_g = 1) = \beta_{0g} + \beta_{1g}D + X^T\beta_{2g} + \varepsilon_g, \qquad g = 1, \ldots, G$$

ovvero, assumendo che il termine di errore ε_g abbia media nulla e varianza costante σ_g^2, e posto $\mu_g = \mu_g(X) = \mathbb{E}(Y|Z_g = 1, D, X)$, diventerà

$$\mu_g = \beta_{0g} + \beta_{1g}D + X^T\beta_{2g}. \tag{3.1}$$

Un primo aspetto da notare riguarda il fatto che, condizionatamente al valore di Z, il modello prevede la possibilità che il trattamento abbia un effetto eterogeneo sulla risposta.

Data l'endogeneità di D, il modello prevede una seconda equazione, che simultaneamente alla (3.1) mette in relazione la probabilità di ricevere il trattamento $\rho_g = \rho_g(X)$ alle variabili di controllo, condizionatamente al valore di Z. Se si considera, ad esempio, un legame logistico, si ha allora:

$$\log\left(\frac{\rho_g}{1-\rho_g}\right) = \delta_{0g} + X^T\delta_{1g} \qquad g = 1, \ldots, G. \tag{3.2}$$

Infine per caratterizzare ulteriormente l'eterogeneità non osservabile, si considerano anche le relazioni tra la probabilità $\pi_g = \pi_g(X)$ di ciascuna modalità di Z e le caratteristiche osservabili X:

$$\log\left(\frac{\pi_g}{\pi_1}\right) = \gamma_{0g} + X^T\gamma_{1g}, \qquad g = 2, \ldots, G. \tag{3.3}$$

Con questo modello si cerca dunque un adattamento ai dati che consideri una caratteristica latente nella popolazione, tenendo conto della relazione tra questa e le caratteristiche X osservate e permettendo che questa porti a differenze in termini di propensione al trattamento e di effetto del trattamento.

3.3 Stima dei parametri

3.3.1 I modelli distributivi di riferimento

I parametri dei modelli (3.1), (3.2), (3.2) vengono stimati simultaneamente mediante il metodo della massima verosimiglianza ed è quindi necessaria la specificazione dei modelli distributivi di riferimento.

Al primo livello si porrà

$$Y|Z_g = 1, D = d, X = x \sim p_g(y|d, x, \theta_g),$$

dove $\theta_g = (\beta_{0g}, \beta_{1g}, \beta_{2g}, \sigma_g^2)$ è il vettore di parametri di regressione che lega Y alla variabile trattamento e alle variabili di controllo, nella (3.1). Data la specificazione (3.1), la scelta più naturale per tale distribuzione è il modello Normale, ovvero $Y|Z_g = 1, D, X \sim N(\mu_g, \sigma_g^2) = N(\beta_{0g} + \beta_{1g}D + X^T\beta_{2g}, \sigma_g^2)$.
Al secondo livello (3.2), volto a trattare l'endogeneità del trattamento, si porrà:

$$D|Z_g = 1, X = x \sim p_D(d|x, \delta_g).$$

dove $\delta_g = (\delta_{0g}, \delta_{1g})$. Nel caso di trattamento dicotomico si pone quindi $D|Z_g = 1, X \sim Bi(1, \rho_g) = Bi\left(1, \frac{e^{\gamma_{0g}+X^T\gamma_{1g}}}{1+e^{\gamma_{0g}+X^T\gamma_{1g}}}\right)$.
Infine, data la natura categoriale assunta per la variabile latente Z si porrà

$$Z|X = x \sim p_Z(z|x, \gamma)$$

con $\gamma = (\gamma_{02}, \gamma_{12}, \ldots, \gamma_{0G}, \gamma_{1G})$, e $Z|X \sim Bi_G(1, \pi)$ con $\pi = (\pi_1, \ldots, \pi_G)$, $\pi_1 = \frac{1}{1+\sum_{h=2}^{G} e^{\gamma_{0h}+X^T\gamma_{1h}}}$ e $\pi_g = \frac{e^{\gamma_{0g}+X^T\gamma_{1g}}}{1+\sum_{h=2}^{G} e^{\gamma_{0h}+X^T\gamma_{1h}}}$, $g = 2, \ldots, G$.
Vale la pena osservare che marginalizzando la densità di $Y|X$ rispetto a Z si ottiene il l'equazione di un modello mistura a G componenti:

$$\begin{aligned} p(y|d, x) &= \sum_{g=1}^{G} p(y, Z_g = 1|d, x) \\ &= \sum_{g=1}^{G} p(Z_g = 1|x)\, p(y|d, x, Z_g = 1) \\ &= \sum_{g=1}^{G} \pi_g\, p_g(y|d, x, \theta_g), \end{aligned}$$

così da rispecchiare l'eterogeneità presente nella popolazione.
La verosimiglianza dei dati, nelle sue versioni osservata e completa, si otterranno a partire dalle distribuzioni congiunte,

$$\begin{aligned} p(y, d|x) &= \sum_{g=1}^{G} p(y, d, Z_g = 1|x) \\ &= \sum_{g=1}^{G} p(Z_g = 1|x)\, p(y, d|x, Z_g = 1) \\ &= \sum_{g=1}^{G} \pi_g\, p_D(d|x, \delta_g)\, p_g(y|d, x, \theta_g). \end{aligned} \tag{3.4}$$

e, rispettivamente

$$p(y, d, z|x) = p_Z(z|x)p_D(d|z, x)p(y|z, d, x). \tag{3.5}$$

3.3.2 L'algoritmo EM

Siano (y_i, d_i, x_i), $i = 1, \ldots, n$, realizzazioni indipendenti dal modello (3.4). Analogamente a quanto fatto nel paragrafo 2.2.1 per modelli mistura di regressione classici, la massimizzazione della verosimiglianza avviene attraverso l'algoritmo EM che richiede la definizione delle verosimiglianze osservata e completa.

A partire dai dati osservati e il modello (3.4) si definisce la funzione di verosimiglianza osservata per i parametri (γ, δ, θ) come

$$\begin{aligned} L^O(\gamma, \delta, \theta|y, d, x) &= \prod_{i=1}^{n} p(y_i, d_i|x_i) \\ &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \, p_D(d_i|x_i, \delta_g) \, p_g(y_i|d_i, x_i, \theta_g), \end{aligned} \tag{3.6}$$

con rispettiva log-verosimiglianza osservata definita come $\ell^O(\gamma, \delta, \theta|y, d, x) = log\left[L^O(\gamma, \delta, \theta|y, d, x)\right]$.

La funzione di verosimiglianza completa, invece, è definita a partire dalla (3.5) assumendo di conoscere, per ogni soggetto, anche il vettore di variabili latenti z_i:

$$\begin{aligned} L^C(\gamma, \delta, \theta|y, d, x, z) &= \prod_{i=1}^{n} p(y_i, d_i, z_i|x_i) = \prod_{i=1}^{n} p(z_i|x_i) \, p_D(d_i|z_i, x_i) p(y_i|d_i, z_i, x_i) \\ &= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g^{z_{ig}}\right] \prod_{g=1}^{G} \left[p_D(d_i|x_i, \delta_g)\right]^{z_{ig}} \prod_{g=1}^{G} \left[p_g(y_i|d_i, x_i, \theta_g)\right]^{z_{ig}} \\ &= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g p_D(d_i|x_i, \delta_g) p_g(y_i|d_i, x_i, \theta_g)\right]^{z_{ig}}, \end{aligned}$$

con rispettiva log-verosimiglianza completa:

$$\begin{aligned} \ell^C(\gamma, \delta, \theta|y, d, x, z) &= \sum_{i=1}^{n} \sum_{g=1}^{G} \log \left[\pi_g p_D(d_i|x_i, \delta_g) p_g(y_i|d_i, x_i, \theta_g)\right]^{z_{ig}} \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \left[\pi_g p_D(d_i|x_i, \delta_g) p_g(y_i|d_i, x_i, \theta_g)\right]. \end{aligned} \tag{3.7}$$

La stima dei parametri si ottiene, analogamente a quanto presentato nel paragrafo 2.2.2, attraverso la massimizzazione del valore atteso della logverosimiglianza completa:

$$\begin{aligned} Q(\gamma, \delta, \theta) &= \mathbb{E}_Z \left[\ell^C(\gamma, \delta, \theta | y, d, x) \right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}_Z \left[Z_{ig} | y, d, x \right] \log \left[\pi_g \, p_D(d_i | x_i, \delta_g) \, p_g(y_i | d_i, x_i, \theta_g) \right], \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig} \left[\log \pi_g + \log p_D(d_i | x_i, \delta_g) + \log p_g(y_i | d_i, x_i, \theta_g) \right], \quad (3.8) \end{aligned}$$

L'algoritmo persegue la massimizzazione attraverso l'alternanza dei passi E, ed M. All' iterazione $(k+1)$-esimo si avrà allora:

E-step Fissato il valore dei parametri $(\gamma^{(k)}, \delta^{(k)}, \theta^{(k)})$ (e quindi dei $\pi_g^{(k)}, \rho_g^{(k)}$ secondo le (3.3), (3.2)) e i dati osservati, l'E-step consiste nel calcolo della (3.8):

$$\begin{aligned} \tau_{ig}^{(k+1)} &= \mathbb{E} \left[Z_{ig} | y, d, x \right] = P(Z_{ig} = 1 | y, d, x) \\ &= \frac{p(Z_{ig} = 1 | x_i) \, p_D(d_i | Z_{ig} = 1, x_i, \delta_g^{(k)}) \, p_g(y_i | Z_{ig} = 1, d_i, x_i, \theta_g^{(k)})}{\sum_{g=1}^{G} p(Z_{ig} = 1 | x_i) \, p_D(d_i | Z_{ig} = 1, x_i, \delta_g^{(k)}) \, p_g(y_i | Z_{ig} = 1, d_i, x_i, \theta_g^{(k)})} \\ &= \frac{\pi_g^{(k)} \, \rho_g^{(k)} \, p_g(y_i | Z_{ig} = 1, d_i, x_i, \theta_g^{(k)})}{\sum_{g=1}^{G} \pi_g^{(k)} \, \rho_g^{(k)} \, p_g(y_i | Z_{ig} = 1, d_i, x_i, \theta_g^{(k)})}, \end{aligned}$$

per $g = 1, \ldots, G$ e per $i = 1, \ldots, n$.

Si noti che la presenza delle probabilità ρ_g nel calcolo delle probabilità a posteriori τ_{ig} mostra come la specificazione del secondo livello (3.2) nel modello per descrivere l'assegnazione del trattamento condizionatamente alle variabili di controllo e a Z aiuti nell'identificazione dell'eterogeneità latente.

M-step Fissate le probabilità a posteriori $\tau_{ig}^{(k+1)}$ nell'E-step, al passo M si determinano i valori aggiornati dei parametri che massimizzano la funzione (3.8)

$$\begin{aligned} (\gamma^{(k+1)}, \delta^{(k+1)}, \theta^{(k+1)}) &= \underset{(\gamma, \delta, \theta)}{\operatorname{argmax}} \, Q(\gamma, \delta, \theta; \tau_{ig}^{(k+1)}) \\ = \underset{(\gamma, \delta, \theta)}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} &\left[\log \pi_g + \log p_D(d_i | x_i, \delta_g) + \log p_g(y_i | d_i, x_i, \theta_g) \right] \end{aligned} \quad (3.9)$$

È utile notare come, a dispetto della maggior complessità del modello, che prevede rispetto ad una semplice mistura di regressione l'inserimento dell'equazione aggiuntiva per spiegare la relazione tra D e le variabili di controllo entro ciascun gruppo, la funzione $Q(\cdot)$ è ancora a parametri separabili, pertanto gli aggiornamenti dei valori possono essere eseguiti distintamente:

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log p_g(y_i | d_i, x_i, \theta_g)$$

$$\delta^{(k+1)} = \underset{\delta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log p_D(d_i | x_i, \delta_g)$$

$$\gamma^{(k+1)} = \underset{\gamma}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(k+1)} \log \pi_g$$

Il procedimento di stima dei parametri dipende dalla scelta delle distribuzioni di riferimento.

Riguardo al parametro θ relativo alla distribuzione della risposta, nel caso particolare illustrato al paragrafo 3.2, di variabile risposta Normale, $\theta_g = (\mu_g, \sigma_g^2) = (\beta_{0g}, \beta_{1g}, \beta_{2g}, \sigma_g^2)$, si può mostrare che l'aggiornamento delle stime ha soluzione esplicita via minimi quadrati pesati (Faria e Soromenho, 2010), per cui si ha che, per $\beta_g = (\beta_{0g}, \beta_{1g}, \beta_{2g})$:

$$\beta_g^{(k+1)} = (X^T W_g^{(k+1)} X)^{-1} X^T W_g^{(k+1)} y, \quad g = 1, \ldots, G$$

dove $W_g^{(k+1)}$ è una matrice diagonale riportante i pesi corrispondenti ai valori $\tau_{ig}^{(k+1)}$ per $i = 1, \ldots, n$; per quanto riguarda l'aggiornamento di σ_g invece:

$$\sigma_g^{2(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ig}^{(k+1)} (y_i - X\beta_g^{(k+1)})}{\sum_{i=1}^{n} \tau_{ig}^{(k+1)}}, \quad g = 1, \ldots, G.$$

I parametri δ del secondo livello della mistura, che considera come risposta l'assegnazione al trattamento regredita sui controlli osservati, hanno un passo di aggiornamento che può essere più o meno computazionalmente oneroso a seconda della distribuzione che viene assunta per D. Con riferimento al trattamento binario specificato nel paragrafo 3.2, o più in generale quando D ha distribuzione appartenente alla famiglia esponenziale, le stime dei parametri δ non sono in generale esprimibili in forma esplicita. Tuttavia, al netto di modifiche minori dovute alla presenza dei pesi τ_{ig}, la massimizzazione è riconducibile all'uso dell'algoritmo dei minimi quadrati pesati iterati (*iterated weighted least squares*,

IWLS), utilizzato per la stima dei modelli lineari generalizzati. Si veda, ad esempio Azzalini (2001).

Nel caso più semplice di trattamento distribuito normalmente, la soluzione di stima si ottiene, come per θ, con i minimi quadrati pesati.

Per quanto riguarda i pesi della mistura, i parametri γ che governano la relazione tra la variabile latente e le variabili di controllo non hanno soluzione di aggiornamento esplicita e sono riconducibili agli algoritmi di stima utilizzati per i modelli di regressione multinomiale. Si vedano, ad esempio, Hasan *et al.* (2014).

E-step ed M-step si alternano fino a convergenza dell'algoritmo.

Come per tutti i modelli mistura, l'algoritmo EM assicura il raggiungimento di un massimo locale e non globale della funzione $Q(\cdot)$, pertanto restano valide le considerazioni fatte sull'importanza dell'inizializzazione dell'algoritmo in conclusione al paragrafo 2.2.2.

Per quanto riguarda la valutazione della variabilità delle stime, è direttamente estendibile al modello proposto la procedura basata sulla matrice d'informazione osservata presentata al paragrafo 2.2.3 con le corrispondenti quantità di verosimiglianza. Lo stesso vale per i criteri di selezione del modello, per la quale si mantiene anche in questo caso il BIC come criterio principale per la scelta del numero delle componenti della mistura.

3.4 Discussione

Il modello proposto risulta essere dunque uno strumento volto alla stima di un effetto causale mediante l'identificazione, e la contestuale inclusione, di potenziali fonti di eterogeneità non direttamente osservabili, e quindi latenti. Se, infatti, l'omissione di variabili rilevanti preclude l'interpretazione del coefficiente di regressione come ATE in presenza di *selection bias*, e i modelli mistura consentono di identificare eventuale eterogeneità non direttamente osservabile, l'utilizzo del modello (3.4) porta ad un vantaggio nella misura in cui si configura come una metodologia volta ad identificare, a partire dai dati, variabili latenti che colgano l'eterogenità e rafforzino la CIA.

L'utilizzo di variabili latenti non è, in realtà, nuovo nell'inferenza causale. I modelli ad equazioni strutturali (*structural equation models*, SEM, Kline e Little, 2011), ad esempio, utilizzati per analizzare le relazioni complesse tra più variabili in un sistema, sono spesso utilizzati per studiare le interazioni e relazioni causali tra variabili latenti (che non possono essere misurate direttamente ma sono invece dedotte da un insieme di variabili osservabili correlate) e variabili osservabili. Un altro esempio recente dell'utilizzo di variabili latenti per l'inferenza causale è quello di Zorzetto *et al.* (2023), dove si introduce un modello mistura bayesiano per individuare gruppi di eterogeneità tra gli individui che

hanno un diverso grado di vulnerabilità rispetto all'inquinamento atmosferico.

Un vantaggio che caratterizza il modello proposto è la sua flessibilità, sia da un punto di vista metodologico che interpretativo. Dal punto di vista metodologico, il modello, illustrato in dettaglio con riferimento a modelli lineari normali e trattamento dicotomico, ricopre in realtà una ben più vasta gamma di situazioni, in quanto immediatamente estendibile all'uso di modelli lineari generalizzati e trattamento appartenente alla famiglia di dispersione esponenziale. Seguendo la notazione classica (si veda, ad esempio, Azzalini, 2001) si avrà allora:

$$\begin{cases} Y|Z_g = 1, D = d, X = x \sim Ep_g(b_Y(\theta_g^{(Y)}), \psi_g) & \text{con probabilità } \pi_g \\ D|Z_g = 1, X = x \sim Ep_g(b_D(\theta_g^{(D)}), \psi_g^{(D)}) & \text{con probabilità } \pi_g(x) \\ Z|X = x \sim Bi_G(1, \pi) \end{cases}$$

dove $T \sim Ep(b_T(\theta^{(T)}), \psi^{(T)})$ se appartiene a una famiglia esponenziale di parametro naturale $\theta^{(T)}$, parametro di dispersione $\psi^{(T)}$ e b_T una funzione nota di θ tale che $\mathbb{E}(T) = b'_T(\theta^{(T)})$. Posto allora $\mu = b'_Y(\theta^{(Y)})$ e $\rho = b'_D(\theta^{(D)})$, i predittori lineari assumeranno la seguente forma:

$$\begin{cases} g_Y(\mu_g) = \beta_{0g} + \beta_{1g} D + X^T \beta_{2g} & g = 1, \ldots, G \\ g_D(\rho_g) = \delta_{0g} + X^T \delta_{1g} & g = 1, \ldots, G \\ \log \frac{\pi_g}{\pi_1} = \gamma_{0g} + x^T \gamma_{1g} & g = 2, \ldots, G. \end{cases}$$

con g_T una opportuna funzione legame che esprime la relazione tra T e le variabili esplicative. I parametri saranno stimati con la massima verosimiglianza mediante l'algoritmo EM che tuttavia, in questa formulazione, includerà a sua volta un'applicazione del metodo dei minimi quadrati pesati e iterati in ciascuna iterazione del passo M.

Con alcune complicazioni di natura principalmente computazionale, inoltre, è possibile aumentare ulteriormente la flessibilità attraverso una modellazione non lineare e l'introduzione di metodi non parametrici, o l'uso di tecniche recenti di selezione delle variabili in contesti di alta dimensionalità dei dati a disposizione.

Dal punto di vista interpretativo, invece, il modello permette di stimare, e testare, anche un'eventuale eterogeneità nell'effetto, non essendo specificato alcun vincolo di uguaglianza degli effetti tra i diversi gruppi identificati. Come anticipato nel paragrafo 1.3 quest'ultimo aspetto è, soprattutto recentemente, di forte interesse nell'ambito della ricerca econometrica, con l'adattamento di metodi di *statistical learning* come ad esempio gli alberi e le foreste causali (Athey e Imbens, 2016; Wager e Athey, 2017), il *double-machine learning* (Chernozhukov *et al.*, 2023), o più recentemente, il *causal clustering* (Kim, 2020, Cap. 4). Quest'ultimo filone, seppur segua logiche anche molto diverse,

presenta una forte analogia con gli strumenti utilizzati in questa tesi, poiché è volto a stimare un effetto causale attraverso l'uso congiunto di metodi classici della letteratura econometrica con tecniche di *clustering*.

Si noti infine che, se l'applicazione naturale della metodologia proposta riguarda gli studi osservazionali, dove è lecito attendersi la presenza di un'eterogeneità che determina un *selection bias*, un ulteriore aspetto di interesse riguarda l'opportunità, in linea di principio, di utilizzare il modello proposto anche all'interno di un esperimento randomizzato o nelle situazioni in cui l'assunzione di *CIA* è da ritenersi soddisfatta. Infatti, il modello proposto include come caso particolare (quando $G = 1$) un modello di regressione classico caratterizzato da trattamento endogeno e pertanto, in linea teorica, non può condurre a peggioramenti in termini di distorsione della valutazione dell'effetto rispetto a quello che si otterrebbe con una stima OLS classica sotto ipotesi di indipendenza condizionale.

Ci sono tuttavia alcune criticità che possono sorgere nell'utilizzo pratico del modello proposto. Il primo aspetto riguarda la natura dei gruppi che la mistura individua: se, in linea teorica, il modello è programmato per individuare gruppi corrispondenti ad un'eterogeneità latente nella popolazione pre-esistente al trattamento, non si può escludere che nella pratica il modello possa individuare una struttura a gruppi configuratasi post-trattamento, come effetto del trattamento stesso. Questa situazione può essere riconducibile al problema noto in letteratura come *bad control* (Cinelli *et al.*, 2020). Aggiungere controllo sull'eterogeneità tra soggetti non porta sempre a risultati migliori in termini di stima di un effetto causale. I cosiddetti *bad control* sono variabili che sono esse stesse un risultato del trattamento, formalizzabili anch'esse attraverso i loro risultati potenziali, che potrebbero essere utilizzate a loro volta come variabile d'interesse o *outcome* del trattamento. La loro inclusione sotto forma di variabili di controllo in un modello che opera sotto CIA induce una nuova fonte di *selection bias* che distorce la stima dell'effetto causale (per una formalizzazione di questa distorsione si veda Angrist e Pischke, 2008). Possono essere tuttavia identificate variabili che sono un effetto del trattamento ma che comunque spiegano in parte una fonte di eterogeneità latente. Queste variabili sono dette *proxy control* e anch'esse inducono distorsione nella stima dell'effetto del trattamento. Ciononostante l'inclusione di una *proxy control* porta a risultati migliori rispetto a non inserire affatto il controllo, pertanto ci si attende che anche in tale situazione l'utilizzo di un modello mistura conduca ad un possibile miglioramento rispetto ad un più semplice modello di regressione con variabili di controllo. Per affrontare questo potenziale problema d'implementazione del modello si suggerisce di porre attenzione all'interpretazione dei gruppi che vengono individuati, riconducendoli possibilmente ad una qualche caratteristica pre-esistente al trattamento, aiutati dalle conoscenze specifiche del contesto analizzato.

Un altro aspetto limitante circa l'implementazione del modello proposto riguarda la natura di caratteristiche pre-esistenti che gli è permesso identificare: per sua costruzione, il modello è in grado di individuare correttamente caratteristiche latenti che possono influenzare la propensione al trattamento ma che presentano soggetti trattati e non trattati per ogni valore che possono assumere. Nel caso in cui la caratteristica latente che si vuole trovare avesse dei valori per cui i soggetti sono solo trattati o solo non trattati, non sarebbe possibile per il modello individuare correttamente l'intera caratteristica portando potenzialmente a risultati non corretti.

Una raccomandazione generale per l'utilizzo di questo strumento è di conoscere l'ambito di applicazione da cui provengono i dati che si stanno analizzando, in modo da poter avere spirito critico nella formulazione delle ipotesi di ciò che si vorrebbe trovare a priori e nell'interpretare correttamente i risultati una volta ottenuti.

Capitolo 4

Uno studio di simulazione

In questo capitolo si presentano i risultati di uno studio empirico sviluppato allo scopo di valutare se, in che misura, e sotto quali condizioni l'individuazione nei dati di gruppi eterogenei di soggetti possa aiutare ad ottenere una corretta quantificazione dell'effetto di un trattamento.

4.1 Valutazione dell'effetto del trattamento

4.1.1 Obiettivi e contesto di lavoro

Il primo passo per valutare l'efficacia e i limiti della metodologia proposta è stato quello di progettare uno studio di simulazione facendo variare alcune condizioni potenzialmente rilevanti. In particolare, sono state considerate le seguenti fonti di variabilità:

- livello di eterogeneità presente nella popolazione, espressa in termini di numero di gruppi e livello di separazione dei gruppi;
- trattamento ricevuto, espresso in termini di numero di trattamenti possibili e probabilità di sottoporvisi;
- effetti omogenei o eterogenei del trattamento tra i soggetti al variare dei gruppi.

Per limitare il numero di confronti da svolgere, gli scenari considerati non prevedono l'inserimento di variabili di controllo. In altre parole, contrariamente al più generale contesto definito nel capitolo precedente, in cui si assumeva che parte dell'eterogeneità della popolazione fosse osservabile, nelle analisi che seguono tutte le fonti di eterogeneità sono da considerarsi latenti.

Riprendendo la notazione introdotta nei capitoli precedenti, nel seguito si assume che ciascun individuo $i, i = 1, \ldots, n$, sia soggetto ad un trattamento D_i del quale si vuole misurare l'effetto con riferimento ad una variabile di interesse $Y_i \in \mathbb{R}$. Ciascun individuo appartiene inoltre ad un gruppo $g, g = 1, \ldots, G$,

definito da una caratteristica non osservata Z_i. La struttura del processo generatore dei dati comune a tutti gli scenari è del tipo:

- $Z_i = (Z_{i1}, \ldots, Z_{iG}) \sim \ Bi_G(1, \pi)$,
  con $\pi = (\pi_1, \ldots, \pi_G)$ e $\pi_g = P(Z_{ig} = 1)$, $g = 1, \ldots, G$ e
  $Z_{ig} = \begin{cases} 1 & \text{se i } \in g \\ 0 & \text{altrimenti} \end{cases}$
- $D_i|(Z_{ig} = 1) = (D_{i0}, D_{i1}, \ldots, D_{iJ})|(Z_{ig} = 1) \sim \ Bi_{J+1}(1, \rho_g)$, con $\rho_g = (\rho_{1g}, \ldots, \rho_{Jg}), g = 1, \ldots, G, \ \rho_{jg} = P(D_{ij} = 1|Z_{ig} = 1)$ e
  $D_{ij} = \begin{cases} 1 & \text{se il soggetto } i \text{ ha ricevuto il trattamento j,} \quad j = 1, \ldots, J \\ 0 & \text{altrimenti} \end{cases}$.
  Si pone, per convenzione, $D_{i0} = 1$ se l'individuo non ha ricevuto alcun trattamento, ovvero se $D_{ij} = 0$ per $j = 1, \ldots, J$.
- $Y_i|(D_i, Z_i) \sim \sum_{g=1}^{G} N(\mu_{ig}, \sigma_g^2) \ I(z_{ig} = 1)$,
  $\mu_{ig} = \beta_{0g} + \sum_{j=1}^{J} \beta_{jg} d_{ij}$,
- $Y_i|(Z_{ig} = 1, D_i = d_i) \sim N(\mu_{ig}, \sigma_g^2)$,
  $\mu_{ig} = \beta_{0g} + D_i' \beta_{1g}$, per $g = 1, \ldots, G$.
  Per semplificare la successiva interpretazione dei risultati, riconducendosi all'analisi di un singolo parametro per ciascuna classe, la $j-$ sima componente del vettore a J componenti β_{1g} sarà posta pari a $j\beta_g$, trattando in questo modo il trattamento alla stregua di una variabile categoriale ordinale con effetto lineare sull'*outcome*.

4.1.2 Dettagli di implementazione

Al fine di avere una discreta gamma di contesti per i quali rispondere alle domanda di interesse, si definiscono le seguenti possibili situazioni:

- L'eterogenità della popolazione si esprime in due o quattro possibili livelli, ovvero $G \in \{2, 4\}$
- Il trattamento è binario (non trattato/trattato) o definito da tre livelli (nessun trattamento, trattamento 1, trattamento 2), ovvero $J \in \{1, 2\}$
- L'effetto del trattmento è omogeneo o eterogeneo su Y

La combinazione delle suddette condizioni si traduce pertanto nella definizione di 8 macro-scenari di simulazione:

- Scenario 1: la caratteristica non osservata è dicotomica, ovvero $Z_i = \{Z_{i1}, Z_{i2}\}$ $(G = 2)$, il trattamento ha due livelli, $D_i = \{D_{i0}, D_{i1}\}$, e con effetto omogeneo nella popolazione $\beta_{1g} = \beta_1, g = 1, \ldots, G$;
- Scenario 2: la caratteristica non osservata è dicotomica, ovvero $Z_i = \{Z_{i1}, Z_{i2}\}$ $(G = 2)$, il trattamento ha due livelli, $D_i = \{D_{i0}, D_{i1}\}$, e con effetto eterogeneo nella popolazione $\beta_{1g} \neq \beta_{1g'}, g \neq g'$;
- Scenario 3: la caratteristica non osservata ha quattro livelli di eteroge-

neità, ovvero $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$ $(G = 4)$, il trattamento ha due livelli, $D_i = \{D_{i0}, D_{i1}\}$, e con effetto omogeneo nella popolazione $\beta_{1g} = \beta_1, g = 1, \ldots, G$;

- Scenario 4: la caratteristica non osservata ha quattro livelli di eterogeneità, ovvero $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$ $(G = 4)$, il trattamento ha due livelli, $D_i = \{D_{i0}, D_{i1}\}$, e con effetto eterogeneo nella popolazione $\beta_{1g} \neq \beta_{1g'}, g \neq g'$;
- Scenario 5: la caratteristica non osservata è dicotomica, ovvero $Z_i = \{Z_{i1}, Z_{i2}\}$ $(G = 2)$, il trattamento ha tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, con effetto omogeneo nella popolazione $\beta_g = \beta_1, g = 1, \ldots, G$;
- Scenario 6: la caratteristica non osservata è dicotomica, ovvero $Z_i = \{Z_{i1}, Z_{i2}\}$ $(G = 2)$, il trattamento ha tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, con effetto eterogeneo nella popolazione $\beta_g \neq \beta_{g'}, g \neq g'$;
- Scenario 7: la caratteristica non osservata ha quattro livelli di eterogeneità, ovvero $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$ $(G = 4)$, il trattamento ha tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, con effetto omogeneo nella popolazione $\beta_g = \beta_1, g = 1, \ldots, G$;
- Scenario 8: la caratteristica non osservata ha quattro livelli di eterogeneità, ovvero $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$ $(K = 4)$, il trattamento ha tre livelli, $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, con effetto eterogeneo nella popolazione $\beta_g \neq \beta_{g'}, g \neq g'$.

Per ciascuno di questi scenari sono stati analizzati degli ulteriori micro-scenari che si distinguono per i valori assegnati ai parametri di probabilità di sottoporsi al trattamento, $P(D_{ij} = 1|Z_{ig} = 1) = \rho_{jg}$ e in termini di separazione tra i livelli di eterogeneità nella popolazione. Per quanto riguarda i parametri ρ_{jg} sono state utilizzate due configurazioni di valori che permettessero di avere un caso in cui tutti il trattamento è assegnato in maniera sostanzialmente bilanciata nella popolazione e un caso in cui almeno uno dei trattamenti è poco assegnato. La separazione tra gruppi è stata invece espressa regolando le varianze σ_g^2, e individuando quattro diverse configurazioni: *cluster* omoschedastici ben distinguibili, *cluster* eteroschedastici ben distinguibili, *cluster* omoschedastici poco distinguibili, *cluster* eteroschedastici poco distinguibili.

La combinazione di questi aspetti ha dato luogo, per ciascuno degli 8 scenario sopra descritti, ad 8 ulteriori micro-scenari, per un totale di $8^2 = 64$ situazioni considerate. Gli specifici parametri di simulazione sono descritti nell'Appendice A ma si riporta, nelle Figure 4.1-4.8 un'illustrazione del processo generatore dei dati per ciascuna delle situazioni considerate, ed un esempio di campione tratto da ciascuna di esse.

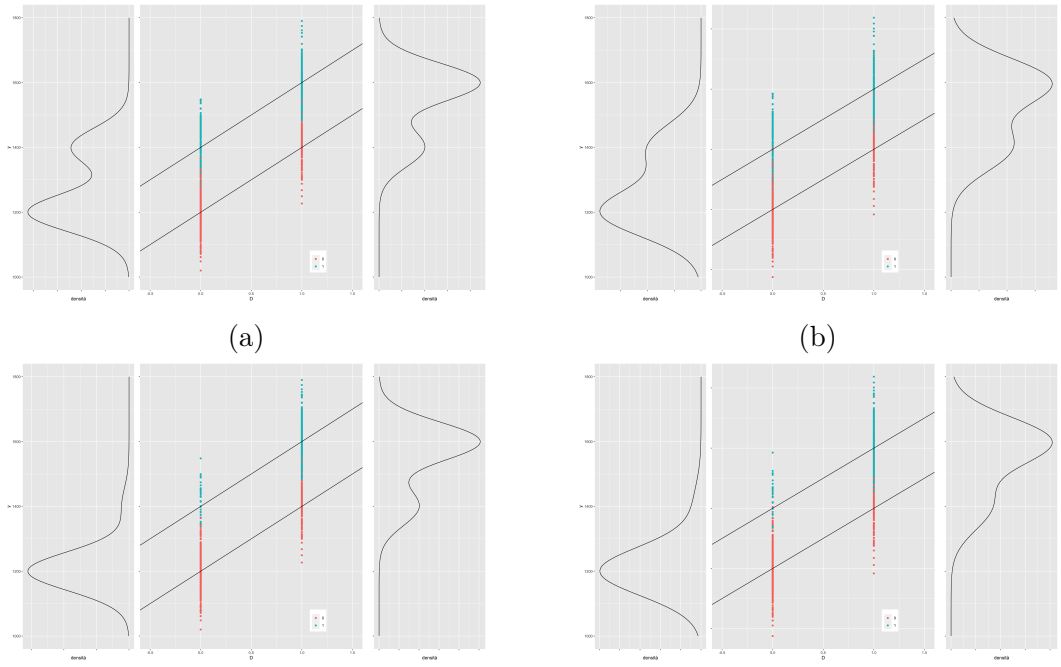
Per ciascuna delle 64 situazioni sopra descritte, sono stati generati 100 campioni di numerosità $n = 1000$. Per ogni campione simulato è stato stimato il modello (3.4) fissando, in questa prima fase, il numero di gruppi pari a quello definito nel modello generatore dei dati. Inoltre, per isolare differenti fonti di

variabilità, l'algoritmo EM è stato impostato con un'inizializzazione dei parametri ragionevolmente vicina ai valori che assumono nel modello generatore dei dati, in modo da ridurre il rischio di convergenza della procedura a massimi locali. Queste ipotesi sono state rilassate nelle elaborazioni riportate nei paragrafi 4.2.1 e 4.2.2.

L'efficacia della procedura proposta è stata valutata:

- in termini assoluti, confrontando la stima dell'effetto con il suo valore noto dal processo generatore dei dati
- in termini relativi, confrontando la stima dell'effetto con quella che si ottiene ignorando le fonti di eterogeneità latente attraverso la semplice applicazione di un modello lineare.

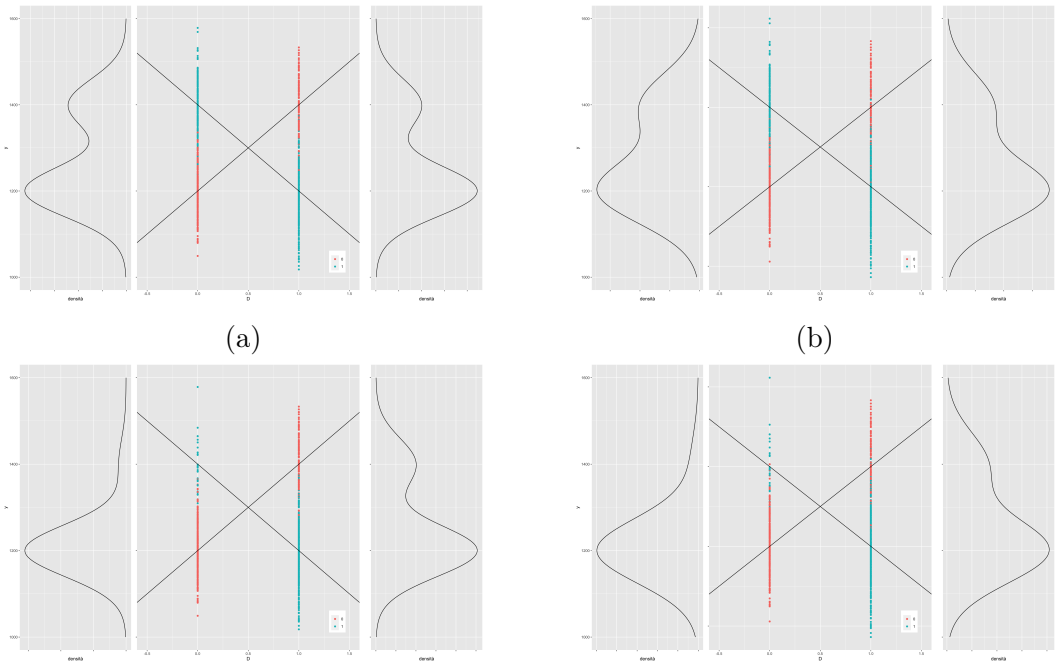
Tutte le analisi sono state svolte nell'ambiente di programmazione R (R Core Team, 2017), utilizzato anche per la predisposizione del codice di implementazione dell'algoritmo EM illustrato al paragrafo 3.3.2.



(a) (b)

(c) (d)

Figura 4.1: Un esempio di campione generato nelle quattro configurazioni omoschedastiche relative allo Scenario 1, modello generatore dei dati e, ai lati, densità teorica di $Y|D$. Le configurazioni di parametri corrispondono a: (a) Assegnazione bilanciata dei trattamenti, cluster omoschedastici e ben distinguibili; (b) Assegnazione bilanciata dei trattamenti, cluster omoschedastici e poco distinguibili; (c) Assegnazione sbilanciata dei trattamenti, cluster omoschedastici e ben distinguibili; (d) Assegnazione sbilanciata dei trattamenti, cluster omoschedastici e poco distinguibili.



(a) (b) (c) (d)

Figura 4.2: Scenario 2. Cfr Figura 4.1.

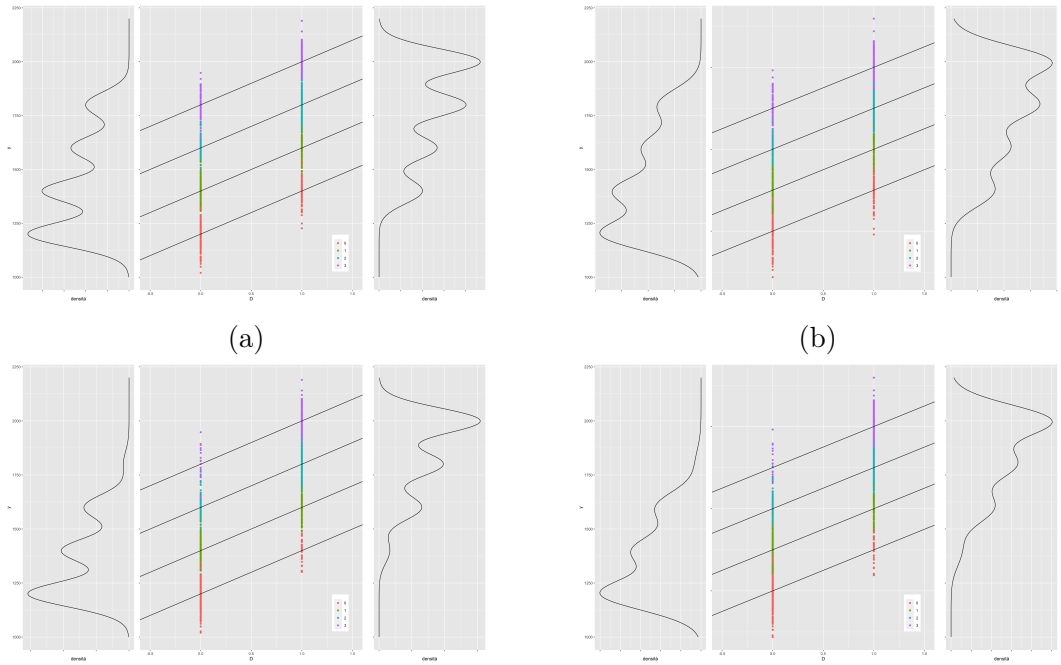
4.1.3 Analisi dei risultati

Nelle Tabelle 4.1-4.8 sono riportati i risultati delle simulazioni descritte, valutati in termini di radice quadrata della media degli errori quadratici medi relativi delle stime rispettivamente dei parametri di intercetta e di coefficiente angolare di gruppo, ovvero:

$$\sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{\beta}_{lg}-\beta_{lg}}{\beta_{lg}}\right)^2}, \quad l=0,1.$$

In ogni Tabella è riportata tra parentesi la stessa misura di errore nel caso in cui si trascuri l'eterogeneità presente nella popolazione adatttando ai dati un modello lineare semplice, ovvero $\beta_{0g}=\beta_0$ e $\beta_{1g}=\beta_1$, $g=1,\ldots,G$.

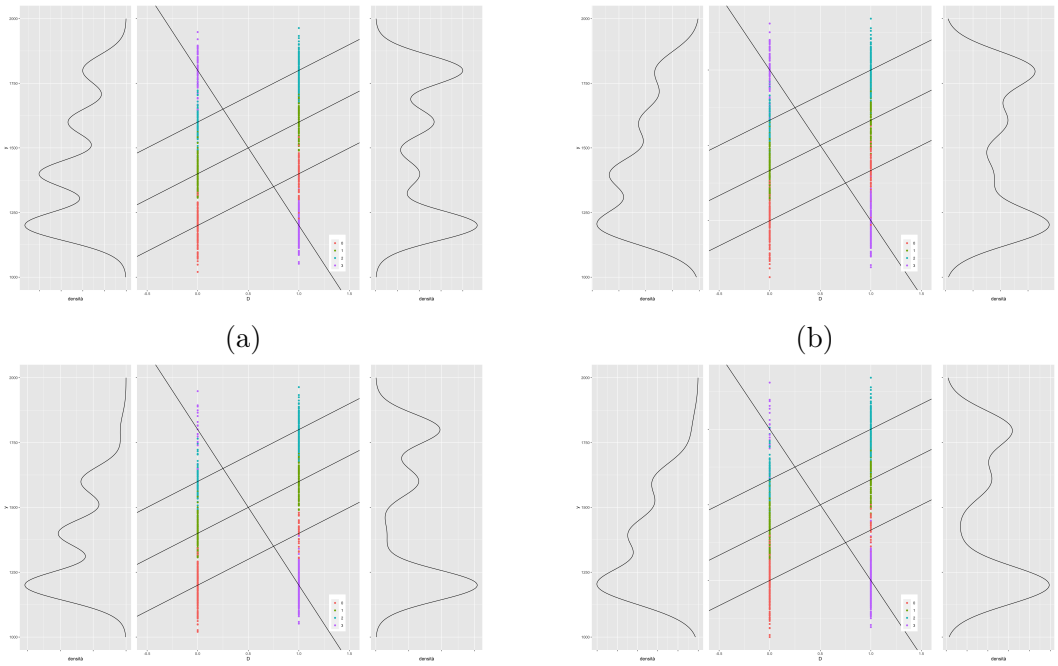
Una prima osservazione generale comune a tutti gli Scenari è che l'errore nelle stime dei parametri diminuisce di molto quando si tiene conto dell'eterogeneità presente nella popolazione, adattando ai dati un modello mistura, rispetto a quando questa viene ignorata e si adatta quindi un modello lineare semplice. Un'altra osservazione comune a tutti gli Scenari è che non si hanno evidenti differenze negli errori di stima tra sotto-scenari che si distinguono esclusivamente per l'omoschedasticità e l'eteroschedasticità dei gruppi, fermo restando il resto delle condizioni. Si nota inoltre che l'errore associato alla stima dell'ef-



(a) (b) (c) (d)

Figura 4.3: Scenario 3. Cfr Figura 4.1.

fetto del trattamento (coefficiente angolare) è sempre maggiore dell'errore che si commette nella stima delle intercette di gruppo. Al variare dei parametri di probabilità di sottoporsi al trattamento e di varianze nei gruppi si osservano invece alcune differenze tra scenari. In generale si può dire che, a partire dalla condizione più favorevole (*cluster* omoschedastici ben distinguibili) fino ad arrivare alla condizione meno favorevole (*cluster* eteroschedastici poco distinguibili), l'errore nelle stime sia delle intercette sia dei coefficienti angolari gradualmente aumenta, e questo aumento risulta essere più accentuato nelle intercette rispetto a quanto si osserva nei coefficienti angolari. Quando si è in presenza di due soli livelli di eterogeneità nella popolazione (Scenari 1,2,5,6) l'errore di stima dei parametri è generalmente più sensibile allo sbilanciamento nell'assegnazione al trattamento rispetto alla situazione in cui si hanno cluster poco distinguibili. Nel caso invece di più livelli di eterogeneità presenti nella popolazione (Scenari 3,4,7,8) si osserva la tendenza opposta: il maggior aumento nell'errore delle stime si ha generalmente con l'aumento delle varianze nei gruppi.



(a) (b) (c) (d)

Figura 4.4: Scenario 4. Cfr Figura 4.1.

4.2 Alcuni aspetti legati alla stima

4.2.1 Identificazione del numero di gruppi eterogenei

È inoltre d'interesse capire se il modello che viene adattato ai dati è in grado di individuare correttamente il numero di gruppi di cui si compone l'eterogeneità della popolazione nei macro e micro-scenari descritti nella sezione 4.1.2. Sono state ripetute le simulazioni assumendo di non conoscere il numero di cluster presenti nei dati, mantenendo sempre un'inizializzazione dei parametri fedele ai valori che assumono nel vero modello generatore. Per ogni campione è quindi stato adattato il metodo di clustering per una sequenza di valori di numero di gruppi ipotizzati ed è stata selezionata la configurazione ottima secondo il BIC.

Nelle Tabelle 4.9-4.16 viene riportato il numero di medio di gruppi che viene selezionato tramite BIC nei campioni simulati per ogni micro-scenario (e il suo errore standard), al variare degli otto macro-scenari. Nei casi in cui sono presenti solo due gruppi di eterogeneità nella popolazione, a prescindere dalla presenza di due o tre livelli di trattamento (Scenari 1,2,5,6), il numero di cluster selezionati dal BIC è praticamente sempre corretto, indipendentemente dalle probabilità di sottoporsi al trattamento nei gruppi e dalle varianze nei gruppi. Quando invece i gruppi di eterogeneità sono più di due (Scenari 3,4,7,8) il nu-

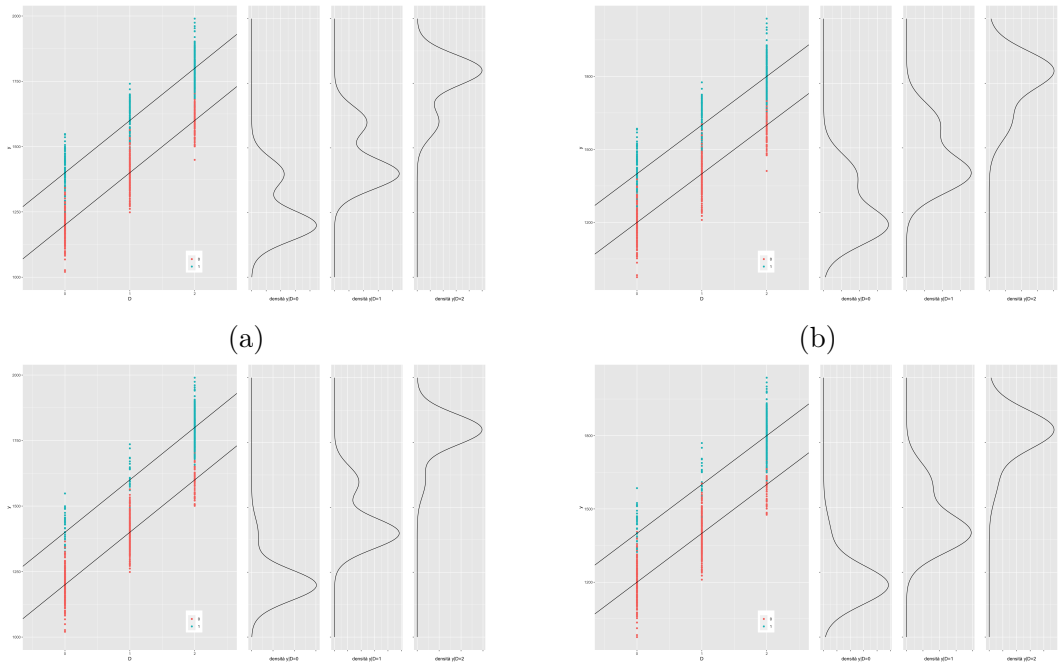


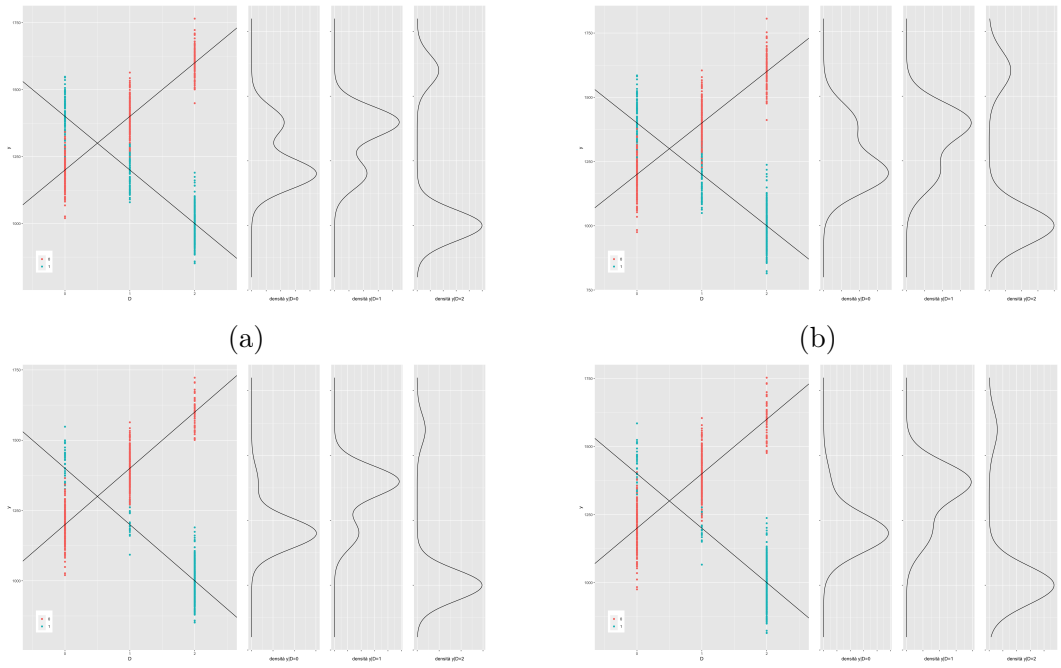
Figura 4.5: Scenario 5. Cfr Figura 4.1.

mero medio di gruppi individuati è vicino al numero di gruppi effettivamente presenti se questi sono ben distinguibili, indipendentemente dal fatto che siano omoschedastici o eteroschedastici e indipendentemente dal fatto che l'assegnazione al trattamento sia bilanciata o meno nei gruppi. Al contrario, quando i gruppi sono poco distinguibili, il BIC tende a selezionare, in media, un numero di gruppi inferiore a quello nominale. Si nota che questo peggioramento è più contenuto nei casi di gruppi con effetti del trattamento eterogenei (Scenari 4 e 8), probabilmente perché differenze così accentuate nell'effetto del trattamento rendono i gruppi tra loro meno aggregabili. In particolare, per lo Scenario 8, si osserva che, in caso di assegnazione sbilanciata dei trattamenti, il numero di gruppi viene identificato quasi sempre correttamente anche in caso di gruppi poco distinguibili; anche in questo caso la questione è attribuibile alla presenza di gruppi con effetti tra loro eterogenei che, nel caso di assegnazione sbilanciata del trattamento, accentuano la loro differenza (si veda Figura 4.8(d)).

4.2.2 Inizializzazione dell'algoritmo

In ultima istanza è stata valutata la sensibilità del metodo all'inizializzazione dei parametri. Per questa analisi ci si è concentrati esclusivamente sugli scenari con effetto del trattamento eterogeneo nella popolazione (Scenari 2-4-6-8).

L'aspetto che rende di particolare interesse questo tipo di analisi è il fatto che

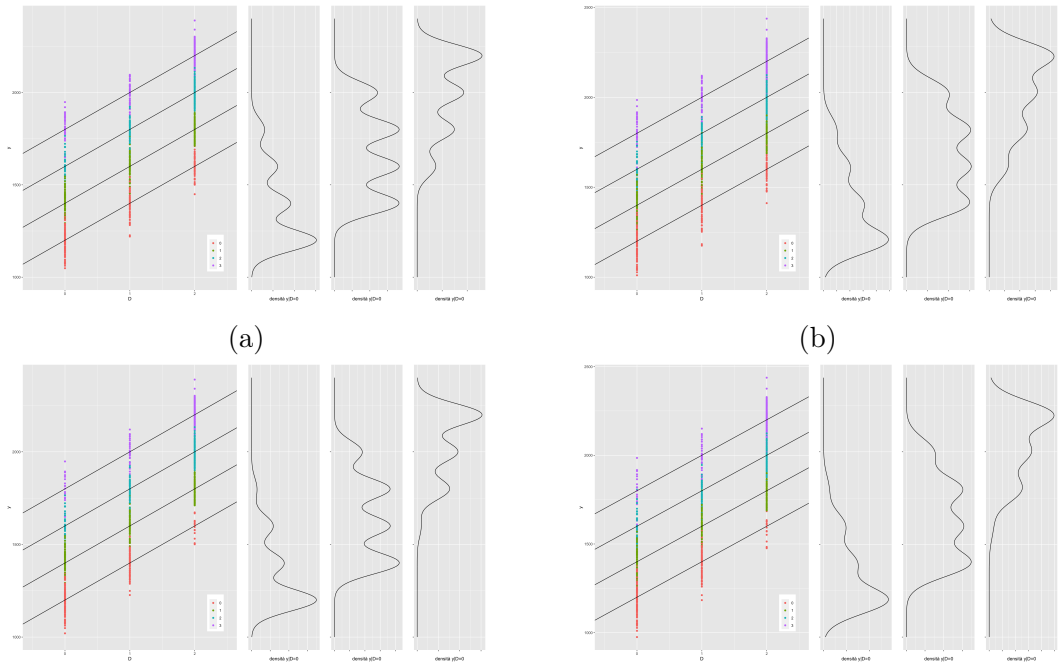


(a) (b) (c) (d)

Figura 4.6: Scenario 6. Cfr Figura 4.1.

le stime sono ottenute tramite algoritmo EM, che non converge necessariamente al massimo globale della verosimiglianza bensì ad un suo massimo locale. Questo rende il metodo potenzialmente sensibile ai valori con cui vengono inizializzati i parametri. La conseguenza che si può avere da un'inizializzazione distante dal vero modello generatore dei dati è di avere una convergenza a stime che comportano una non corretta identificazione dei gruppi di eterogeneità e quindi ad una meno accurata quantificazione degli effetti del trattamento. Per gli Scenari considerati questo si traduce in stime che non colgono a pieno l'eterogeneità che si ha nell'effetto del trattamento tra gruppi.

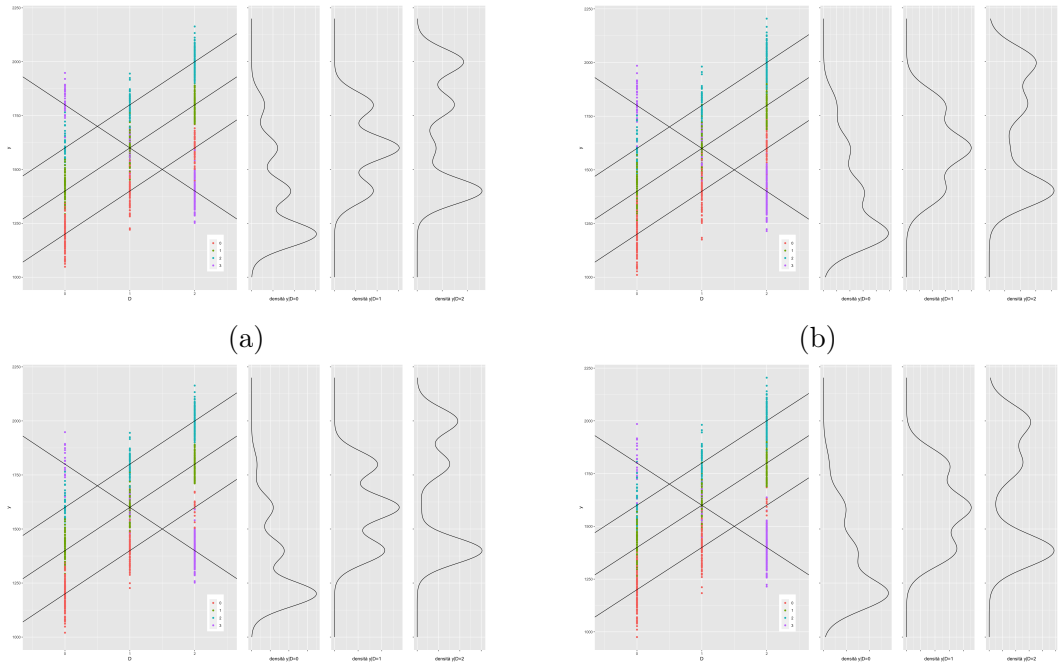
È d'interesse capire se il metodo utilizzato è in grado di riconoscere come massimo globale della verosimiglianza le stime che corrispondono a una corretta identificazione dei gruppi di eterogeneità con i relativi effetti del trattamento. Nel caso in cui questo non si verificasse, è d'interesse invece capire entro quali limiti di inizializzazione il metodo converge al massimo locale corrispondente all'identificazione dei gruppi generati dal vero modello generatore dei dati. Sono state quindi calcolate e confrontate le log-verosimiglianze corrispondenti a convergenze che portano ad identificazioni dei gruppi di eterogeneità corrette e non corrette per ogni micro-scenario dei quattro macro-scenari considerati. Nelle Tabelle 4.17-4.19 sono riportate le log-verosimiglianze medie che si ottengono tra i campioni simulati quando l'algoritmo converge a delle stime che identificano correttamente i cluster presenti e quando converge a delle stime



(a) (b) (c) (d)

Figura 4.7: Scenario 7. Cfr Figura 4.1.

che non identificano i veri cluster. La tabella relativa allo Scenario 6 è assente in quanto non si riscontra alcuna inizializzazione dei parametri che porti ad una convergenza diversa da quella che identifica correttamente i gruppi presenti e i relativi effetti del trattamento; questo è probabilmente dovuto alla particolarità di questo Scenario di simulazione in cui i due unici gruppi presenti sono caratterizzati da effetti del trattamento a tre livelli che portano ad un'ampia distanza tra i gruppi (si veda Figura 4.6). Per gli Scenari che prevedono due livelli di trattamento (Scenari 2 e 4) si può vedere che il principale discriminante tra i casi in cui alle stime che identificano correttamente i cluster corrisponde il massimo globale della verosimiglianza e i casi in cui questo non avviene è il fatto che i cluster presenti siano omoschedastici o eteroschedastici: quando i cluster sono eteroschedastici, infatti, il modello mistura riconosce come più verosimili i gruppi che hanno la stessa varianza tra i livelli di trattamento; se i cluster sono omoschedastici, invece, non si può sfruttare questa informazione e dunque diverse configurazioni degli effetti del trattamento possono essere ritenute equivalentemente verosimil. Fa eccezione a questa osservazione il caso in cui i cluster sono eteroschedastici ma poco distinguibili nello Scenario 4: in questo caso risulta comunque più verosimile l'identificazione non corretta dei cluster e dei corrispondenti effetti del trattamento nella popolazione e ciò è probabilmente dovuto alla configurazione (piuttosto estrema) prevista per i veri effetti del trattamento in questo Scenario (si veda Figura 4.4). Nei casi in



(a) (b) (c) (d)

Figura 4.8: Scenario 8. Cfr Figura 4.1.

cui le stime dei parametri che identificano correttamente i cluster corrispondono solo a massimi locali e non globali in questi due Scenari, l'algoritmo EM converge a stime fedeli al vero modello generatore dei dati solo se i valori iniziali dei parametri corrispondenti agli effetti del trattamento (β_1) danno una chiara indicazione di effetti del trattamento eterogenei (in questo caso di segno opposto) nella popolazione, altrimenti, anche senza dare alcuna indicazione iniziale sugli effetti del trattamento, l'algoritmo EM converge a stime che non corrispondono alle corrette identificazioni dei cluster presenti. Una rappresentazione grafica di quanto accade in questi casi viene riportata in Figura 4.9.

Per lo Scenario 8 si osserva invece che la corretta identificazione dei cluster corrisponde sempre al massimo globale della verosimiglianza; questo può essere spiegato dal fatto che, essendo (come lo Scenario 6) uno Scenario a tre livelli di trattamento con effetto lineare nei livelli, il modello mistura (che modella linearmente l'effetto del trattamento) è aiutato ad identificare correttamente gli effetti del trattamento rispetto al caso in cui i livelli di trattamento sono solo due.

Scenario 1	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0039 (0.0773)	0.0090 (0.0944)
	β_1	0.0337 (0.3037)	0.0670 (0.6905)
Cluster eteroschedastici ben distinguibili	β_0	0.0039 (0.0773)	0.0094 (0.0944)
	β_1	0.0337 (0.3037)	0.0692 (0.6904)
Cluster omoschedastici poco distinguibili	β_0	0.0068 (0.0773)	0.0150 (0.0944)
	β_1	0.0508 (0.3040)	0.1098 (0.6903)
Cluster eteroschedastici poco distinguibili	β_0	0.0069 (0.0773)	0.0154 (0.0944)
	β_1	0.0504 (0.3040)	0.1114 (0.6902)

Tabella 4.1: Radice quadrata della media degli errori quadratici medi relativi delle stime dei parametri β_0 e β_1 al variare dei campioni Monte Carlo. Tra parentesi il medesimo indicatore è calcolato con riferimento alle stime ottenute ignorando l'eterogeneità latente. Risultati relativi allo Scenario 1.

Scenario 2	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0039 (0.0773)	0.0110 (0.0944)
	β_1	0.0441 (1.0011)	0.0853 (1.0149)
Cluster eteroschedastici ben distinguibili	β_0	0.0038 (0.0773)	0.0115 (0.0944)
	β_1	0.0436 (1.0011)	0.0888 (1.0149)
Cluster omoschedastici poco distinguibili	β_0	0.0067 (0.0773)	0.0201 (0.0944)
	β_1	0.0845 (1.0012)	0.1584 (1.0150)
Cluster eteroschedastici poco distinguibili	β_0	0.0066 (0.0773)	0.0206 (0.0944)
	β_1	0.0825 (1.0012)	0.1615 (1.0150)

Tabella 4.2: Cfr. Tab. 4.1. Risultati relativi allo Scenario 2.

Scenario 3	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0063 (0.1508)	0.0115 (0.1591)
	β_1	0.0572 (0.7084)	0.1120 (1.3767)
Cluster eteroschedastici ben distinguibili	β_0	0.0062 (0.1508)	0.0119 (0.1591)
	β_1	0.0572 (0.7082)	0.1139 (1.3767)
Cluster omoschedastici poco distinguibili	β_0	0.0153 (0.1508)	0.0211 (0.1591)
	β_1	0.1114 (0.7084)	0.1952 (1.3764)
Cluster eteroschedastici poco distinguibili	β_0	0.0142 (0.1508)	0.0218 (0.1591)
	β_1	0.1093 (0.7082)	0.1936 (1.3764)

Tabella 4.3: Cfr. Tab. 4.1. Risultati relativi allo Scenario 3.

Scenario 4	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0060 (0.1508)	0.0112 (0.1591)
	β_1	0.0567 (0.8223)	0.0973 (0.7225)
Cluster eteroschedastici ben distinguibili	β_0	0.0059 (0.1508)	0.0115 (0.1591)
	β_1	0.0554 (0.8224)	0.0932 (0.7225)
Cluster omoschedastici poco distinguibili	β_0	0.0135 (0.1508)	0.0175 (0.1591)
	β_1	0.1409 (0.8226)	0.2049 (0.7223)
Cluster eteroschedastici poco distinguibili	β_0	0.0128 (0.1508)	0.0175 (0.1591)
	β_1	0.1352 (0.8226)	0.1912 (0.7223)

Tabella 4.4: Cfr. Tab. 4.1. Risultati relativi allo Scenario 4.

Scenario 5	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0044 (0.0796)	0.0296 (0.0993)
	β_1	0.0211 (0.2141)	0.1039 (0.7225)
Cluster eteroschedastici ben distinguibili	β_0	0.0045 (0.0796)	0.0219 (0.0993)
	β_1	0.0210 (0.2141)	0.0771 (0.7225)
Cluster omoschedastici poco distinguibili	β_0	0.0078 (0.0796)	0.0577 (0.0953)
	β_1	0.0323 (0.2141)	0.2028 (0.7223)
Cluster eteroschedastici poco distinguibili	β_0	0.0079 (0.0796)	0.0544 (0.0931)
	β_1	0.0320 (0.2141)	0.1912 (0.7223)

Tabella 4.5: Cfr. Tab. 4.1. Risultati relativi allo Scenario 5.

Scenario 6	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0038 (0.1508)	0.0289 (0.1591)
	β_1	0.0196 (0.8223)	0.1044 (0.7225)
Cluster eteroschedastici ben distinguibili	β_0	0.0039 (0.1508)	0.0270 (0.1591)
	β_1	0.0194 (0.8224)	0.0956 (0.7225)
Cluster omoschedastici poco distinguibili	β_0	0.0053 (0.1508)	0.0519 (0.1591)
	β_1	0.0263 (0.8226)	0.1840 (0.7223)
Cluster eteroschedastici poco distinguibili	β_0	0.0054 (0.1508)	0.0541 (0.1591)
	β_1	0.0262 (0.8226)	0.1908 (0.7223)

Tabella 4.6: Cfr. Tab. 4.1. Risultati relativi allo Scenario 6.

Scenario 7	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0075 (0.1557)	0.0083 (0.1659)
	β_1	0.0366 (0.5275)	0.0436 (0.7311)
Cluster eteroschedastici ben distinguibili	β_0	0.0075 (0.1557)	0.0082 (0.1659)
	β_1	0.0363 (0.5274)	0.0426 (0.7311)
Cluster omoschedastici poco distinguibili	β_0	0.0194 (0.1557)	0.0259 (0.1659)
	β_1	0.0899 (0.5275)	0.1231 (0.7308)
Cluster eteroschedastici poco distinguibili	β_0	0.0192 (0.1557)	0.0209 (0.1659)
	β_1	0.0877 (0.5274)	0.1068 (0.7308)

Tabella 4.7: Cfr. Tab. 4.1. Risultati relativi allo Scenario 7.

Scenario 8	Parametro	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	β_0	0.0076 (0.1518)	0.0148 (0.1556)
	β_1	0.0390 (0.8924)	0.0756 (0.9090)
Cluster eteroschedastici ben distinguibili	β_0	0.0076 (0.1518)	0.0145 (0.1556)
	β_1	0.0392 (0.8924)	0.0725 (0.9090)
Cluster omoschedastici poco distinguibili	β_0	0.0139 (0.1518)	0.0218 (0.1556)
	β_1	0.0704 (0.8924)	0.1246 (0.9089)
Cluster eteroschedastici poco distinguibili	β_0	0.0139 (0.1518)	0.0203 (0.1556)
	β_1	0.0695 (0.8924)	0.1019 (0.9089)

Tabella 4.8: Cfr. Tab. 4.1. Risultati relativi allo Scenario 8.

Scenario 1	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster eteroschedastici ben distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster omoschedastici poco distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster eteroschedastici poco distinguibili	2.00 (0.00)	2.00 (0.00)

Tabella 4.9: Media (e deviazione standard) del numero di *cluster* individuati al variare dei campioni Monte Carlo. Risultati relativi allo Scenario 1.

Scenario 2	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster eteroschedastici ben distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster omoschedastici poco distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster eteroschedastici poco distinguibili	2.00 (0.00)	2.00 (0.00)

Tabella 4.10: Cfr. Tab. 4.9. Risultati relativi allo Scenario 2.

Scenario 3	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	3.90 (0.30)	3.85 (0.36)
Cluster eteroschedastici ben distinguibili	3.93 (0.26)	3.90 (0.30)
Cluster omoschedastici poco distinguibili	2.85 (0.46)	2.76 (0.51)
Cluster eteroschedastici poco distinguibili	2.85 (0.50)	2.72 (0.55)

Tabella 4.11: Cfr. Tab. 4.9. Risultati relativi allo Scenario 3.

Scenario 4	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	3.78 (0.42)	3.71 (0.46)
Cluster eteroschedastici ben distinguibili	3.80 (0.40)	3.64 (0.48)
Cluster omoschedastici poco distinguibili	3.00 (0.14)	3.00 (0.28)
Cluster eteroschedastici poco distinguibili	3.02 (0.14)	3.03 (0.22)

Tabella 4.12: Cfr. Tab. 4.9. Risultati relativi allo Scenario 4.

Scenario 5	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	2.00 (0.00)	2.06 (0.24)
Cluster eteroschedastici ben distinguibili	2.00 (0.00)	2.03 (0.18)
Cluster omoschedastici poco distinguibili	2.00 (0.00)	2.08 (0.27)
Cluster eteroschedastici poco distinguibili	2.00 (0.00)	2.03 (0.18)

Tabella 4.13: Cfr. Tab. 4.9. Risultati relativi allo Scenario 5.

Scenario 6	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	2.00 (0.00)	2.06 (0.24)
Cluster eteroschedastici ben distinguibili	2.00 (0.00)	2.00 (0.00)
Cluster omoschedastici poco distinguibili	2.00 (0.00)	2.06 (0.24)
Cluster eteroschedastici poco distinguibili	2.00 (0.00)	2.01 (0.10)

Tabella 4.14: Cfr. Tab. 4.9. Risultati relativi allo Scenario 6.

Scenario 7	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	3.85 (0.36)	3.93 (0.26)
Cluster eteroschedastici ben distinguibili	3.91 (0.29)	3.95 (0.22)
Cluster omoschedastici poco distinguibili	2.90 (0.44)	2.96 (0.45)
Cluster eteroschedastici poco distinguibili	2.84 (0.49)	2.97 (0.46)

Tabella 4.15: Cfr. Tab. 4.9. Risultati relativi allo Scenario 7.

Scenario 8	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	3.99 (0.1)	3.98 (0.14)
Cluster eteroschedastici ben distinguibili	3.97 (0.17)	3.99 (0.10)
Cluster omoschedastici poco distinguibili	3.34 (0.48)	3.92 (0.27)
Cluster eteroschedastici poco distinguibili	3.74 (0.44)	3.97 (0.17)

Tabella 4.16: Cfr. Tab. 4.9. Risultati relativi allo Scenario 8.

Scenario 2	Identificazione dei cluster	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	corretta	-6050.28	-5873.03
	non corretta	**-6050.26**	**-5872.76**
Cluster eteroschedastici ben distinguibili	corretta	**-6046.62**	**-5868.60**
	non corretta	-6048.31	-5871.42
Cluster omoschedastici poco distinguibili	corretta	-6179.42	-6032.11
	non corretta	**-6179.36**	**-6031.86**
Cluster eteroschedastici poco distinguibili	corretta	**-6177.09**	**-6028.87**
	non corretta	-6177.46	-6029.90

Tabella 4.17: Log-verosimiglianza media al variare dei campioni Monte Carlo a convergenza dell'algoritmo EM nei casi in cui si ottengano una corretta e una non corretta identificazione dei cluster presenti nella popolazione. In grassetto il valore più alto. Risultati relativi allo Scenario 2

Scenario 4	Identificazione dei cluster	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	corretta	-6679.02	-6534.83
	non corretta	**-6678.26**	**-6533.66**
Cluster eteroschedastici ben distinguibili	corretta	**-6675.26**	**-6530.90**
	non corretta	-6675.79	-6531.40
Cluster omoschedastici poco distinguibili	corretta	-6765.13	-6645.41
	non corretta	**-6764.24**	**-6644.51**
Cluster eteroschedastici poco distinguibili	corretta	-6762.44	-6642.07
	non corretta	**-6761.90**	**-6641.15**

Tabella 4.18: Cfr. Tab. 4.17. Risultati relativi allo Scenario 4

Scenario 8	Identificazione dei cluster	Assegnazione bilanciata dei trattamenti	Assegnazione sbilanciata dei trattamenti
Cluster omoschedastici ben distinguibili	corretta	**-6572.53**	**-6511.10**
	non corretta	-6599.92	-6533.63
Cluster eteroschedastici ben distinguibili	corretta	**-6567.61**	**-6506.39**
	non corretta	-6598.71	-6527.69
Cluster omoschedastici poco distinguibili	corretta	**-6664.19**	**-6618.40**
	non corretta	-6670.81	-6621.73
Cluster eteroschedastici poco distinguibili	corretta	**-6659.95**	**-6614.44**
	non corretta	-6664.74	-6616.89

Tabella 4.19: Cfr. Tab. 4.17. Risultati relativi allo Scenario 8

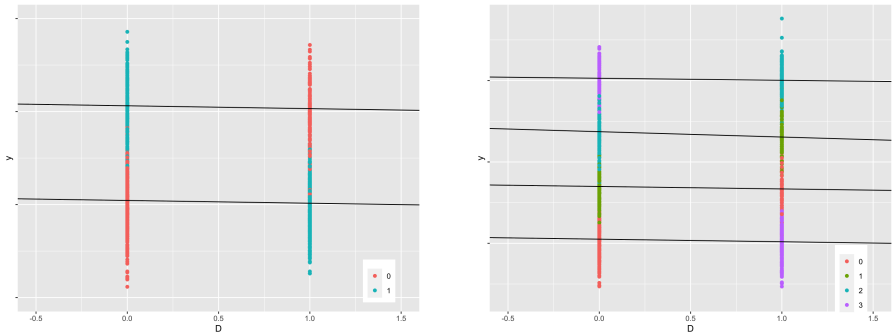


Figura 4.9: Esempio di stime associate a inizializzazioni che non conducono a una corretta valutazione degli effetti del trattamento, per gli Scenari 2 (a sinistra) e 4 (a destra).

Capitolo 5

Un'applicazione a dati reali

In questo Capitolo si presentano e discutono i risultati ottenuti dall'applicazione della metodologia sviluppata nel Capitolo 3 ad un insieme di dati reali, con l'obiettivo di stimare l'effetto degli anni d'istruzione sul reddito di un individuo. I dati utilizzati sono stati precedentemente analizzati con tecniche di regressione con variabili di controllo e variabili strumentali da Brunello *et al.* (2017). I risultati di tale studio saranno pertanto considerati come termini di confronto.

5.1 I dati *SHARE*

5.1.1 Descrizione dei dati

I dati analizzati provengono dal *Survey of Health, Ageing and Retirement in Europe* (*SHARE*), un'indagine, finanziata dalla Commissione Europea, volta a raccogliere informazioni riguardanti diversi ambiti in campo socio-economico per più di 25000 individui a partire dai 50 anni di età. In particolare, è d'interesse il dataset *SHARELIFE*, che contiene variabili su reddito, passato e presente (reale e PPP-adjusted), coorte di nascita, Stato di provenienza e moltre altre caratteristiche d'interesse di un campione di soggetti, tra cui alcune riguardanti la loro infanzia. Le analisi sono state svolte su un campione di $n = 5390$ individui, maschi, nati tra il 1920 e il 1956 e residenti in Austria, Belgio, Repubblica Ceca, Danimarca, Francia, Germania, Italia, Paesi Bassi e Svezia. Seguendo Brunello *et al.* (2017), vengono escluse dall'analisi le lavoratrici donne, per problemi associati alla partecipazione femminile al mercato del lavoro, i lavoratori autonomi, e i lavoratori occupati da meno di 5 anni, e si concentra l'analisi sulle seguenti variabili:

- *Current* (*main*) *wage*: reddito dell'individuo al momento dell'intervista (o reddito del principale lavoro svolto nella propria carriera lavorativa nel caso di individui in pensione) convertito in Euro;

- *Lifetime earnings*: reddito permanente, convertito in Euro (per i dettagli del suo calcolo si veda Brunello *et al.*, 2017);
- *Years of education*: anni d'istruzione;
- *Years of compulsory education*: anni d'istruzione obbligatoria;
- *Cohort*: variabile quantitativa definita come (coorte di nascita - 1919);
- *Rural*: variabile indicatrice che ha valore 1 se il soggetto viveva in un'area rurale all'età di 10 anni, 0 altrimenti;
- *Book*: variabile indicatrice che ha valore 1 se il soggetto viveva con pochi libri (da 0 a 10 libri) non scolastici in casa all'età di 10 anni, 0 altrimenti (11 o più libri);
- *SE*, *DK*, *DEW*, *NL*, *FR*, *AT*, *IT*, *CZ*: variabili indicatrice relative al paese di provenienza; valgono 1 se il soggetto proviene rispettivamente da Svezia, Danimarca, Germania, Paesi Bassi, Francia, Austria, Italia e Repubblica Ceca, 0 altrimenti.

Date le variabili rilevate, l'obiettivo è valutare l'effetto degli anni di istruzione (*Years of education*) sul reddito, alternativamente misurato dalle variabili *Current wage* e *Lifetime earnings*.

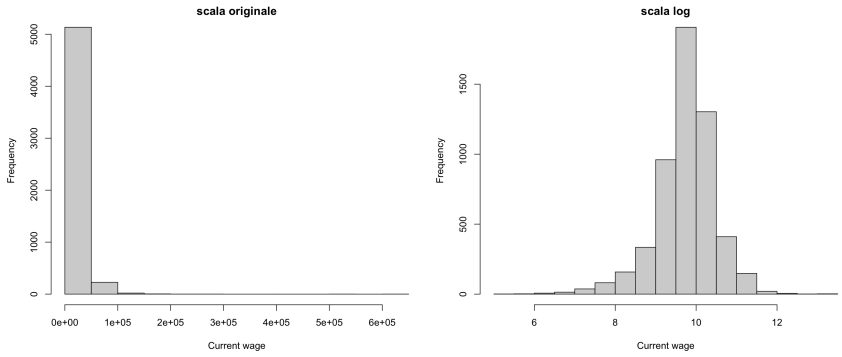
5.1.2 Alcune analisi esplorative

Si conducono alcune analisi esplorative principalmente grafiche sulle variabili disponibili per poter comprenderne la natura e le relazioni, ed orientare la successiva fase di modellazione.

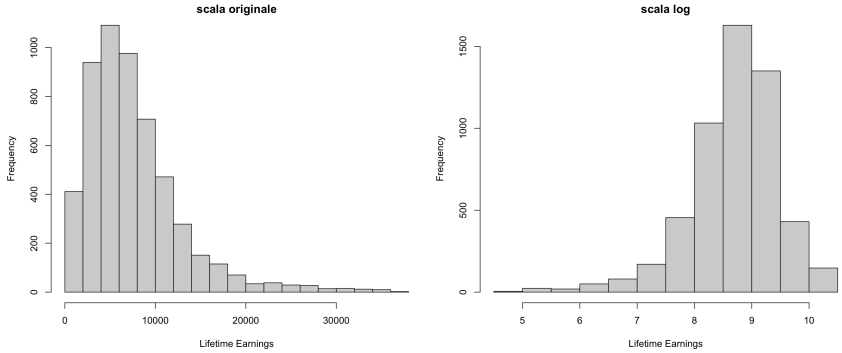
Nella Figura 5.1 sono riportate le distribuzioni marginali delle due possibili variabili risposta nella scala originale e nella scala logaritmica, che ne simmetrizza le forme, in particolare per la variabile *Current wage* e in modo meno efficace per la variabile *Lifetime Earnings*.

Per quanto riguarda la variabile di trattamento, gli anni d'istruzione sono una variabile quantitativa discreta che assume valori da 2 a 25 e con una distribuzione piuttosto simmetrica attorno alla sua media (Figura 5.2). Si nota un picco di frequenza in corrispondenza degli 8 anni d'istruzione, probabilmente dovuto al fatto che questo valore corrisponde nella maggior parte degli Stati al numero di anni d'istruzione obbligatoria.

Con riferimento alle relazioni marginali tra le variabili risposta *Current wage* e *Lifetime Earnings* (in scala logaritmica) e le altre variabili, in entrambi i casi si osserva un'associazione positiva con il trattamento. Al contrario, non si evince alcuna chiara tendenza in relazione alla coorte di nascita e alla provenienza da un'area rurale (*Rural*, Figura 5.3). In merito alla distribuzione nei diversi Stati di provenienza, risaltano in particolare i redditi più alti della media per gli abitanti di Svezia (*SE*) e Germania (*DEW*) (e Danimarca *DK* nel caso di *Lifetime earnings*) e più bassi per Italia (*IT*) e Repubblica Ceca (*CZ*).



(a)



(b)

Figura 5.1: Distribuzioni marginali delle variabili risposta (a) *Current wage* e (b) *Lifetime earnings* in scala originale (sinistra) e in scala logaritmica (destra).

5.2 Stima dell'effetto dell'istruzione sul reddito

5.2.1 I modelli adottati

L'obiettivo dell'analisi è indagare se, e in che misura, la metodologia proposta nel Capitolo 3 possa essere uno strumento utile per la stima dell'effetto d'interesse e l'identificazione di eventuali *pattern* di eterogeneità.

Considerata la natura delle variabili, appare ragionevole assumere una distribuzione Normale sia per descrivere il modello per l'*outcome*, su scala logaritmica, che la variabile di trattamento.

Il modello mistura a G componenti è dunque specificato come segue:

$$\begin{cases} \log(Y_i)|Z_{ig}=1, D_i, X_i \sim N(\mu_g, \sigma_g^2) & \text{con probabilità } \pi_g \\ D_i|Z_{ig}=1, X_i \sim N(\nu_g, \psi_g^2) & \text{con probabilità } \pi_g \\ Z_i|X_i \sim Bi_G(1,\pi) & i=1,\dots,n. \end{cases}$$

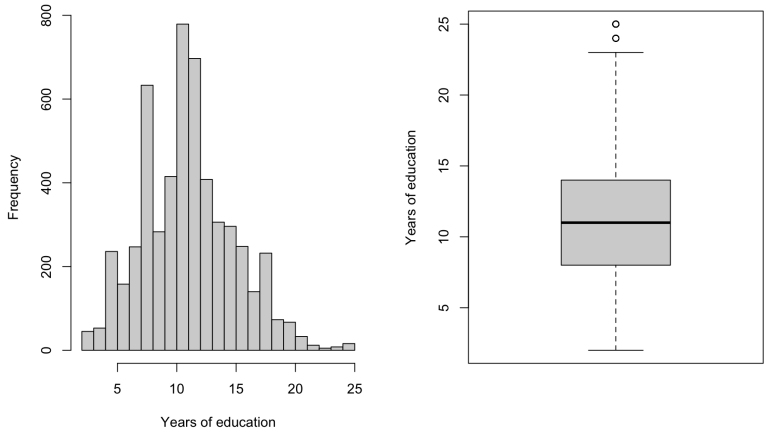


Figura 5.2: Distribuzione marginale del trattamento *Years of education*.

dove

$$\begin{cases} \mu_g = \beta_{0g} + \beta_{1g} D_i + X_i^T \beta_{2g} & \text{con probabilità } \pi_g \\ \nu_g = \delta_{0g} + X_i^T \delta_{1g} & \text{con probabilità } \pi_g \\ \log\left(\frac{\pi_g}{\pi_1}\right) = X_i^T \gamma_{1g} & i = 1, \ldots, n. \end{cases} \tag{5.1}$$

e $X_i = (Cohort, Rural, SE, DK, DEW, NL, FR, AT, IT, CZ)^T$.

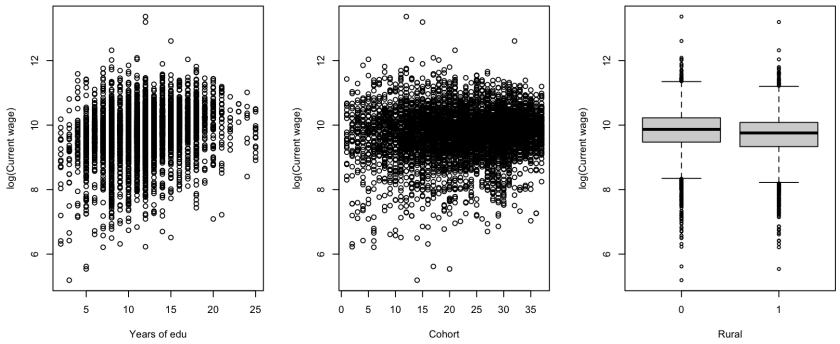
I parametri del modello vengono stimati, al variare del numero G delle classi latenti considerate, via massima verosimiglianza, applicando l'algoritmo EM descritto al paragrafo 3.3.2.

I risultati, riportati nel paragrafo che segue, vengono messi a confronto con il lavoro di Brunello *et al.* (2017). Per la valutazione dell'effetto, gli autori prendono a riferimento un modello lineare stimato con i minimi quadrati ordinari (OLS) con inclusione di variabili controllo:

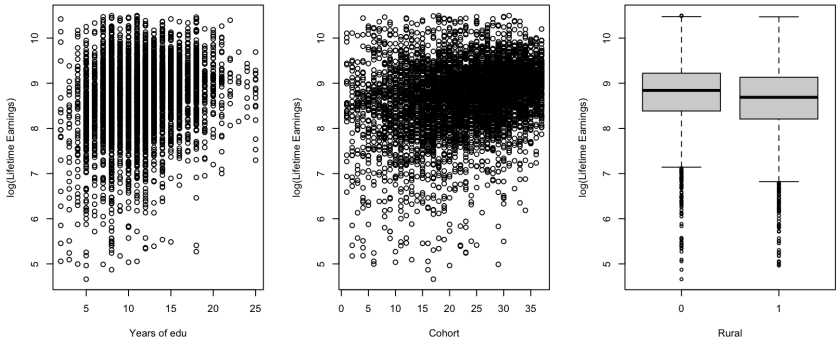
$$\log(Y_i) = \beta_0 + \beta_1 D_i + X_i^T \beta_2 + \varepsilon_i \quad i = 1, \ldots, n$$

e propongono un modello a due stadi (*two-stages least squares*, 2SLS) che utilizza come variabili strumentali, contenute nel vettore S_i, gli anni d'istruzione obbligatoria e la loro interazione con la variabile *Rural*, poiché le riforme sull'istruzione obbligatoria nei diversi Stati sono state maggiormente influenti nelle zone rurali, dove la partecipazione all'istruzione era meno saliente prima della loro introduzione:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + X_i^T \beta_2 + U_i \\ D_i &= \alpha_0 + S_i^T \alpha_1 + X_i^T \alpha_2 + V_i \quad i = 1, \ldots, n. \end{aligned}$$



(a)



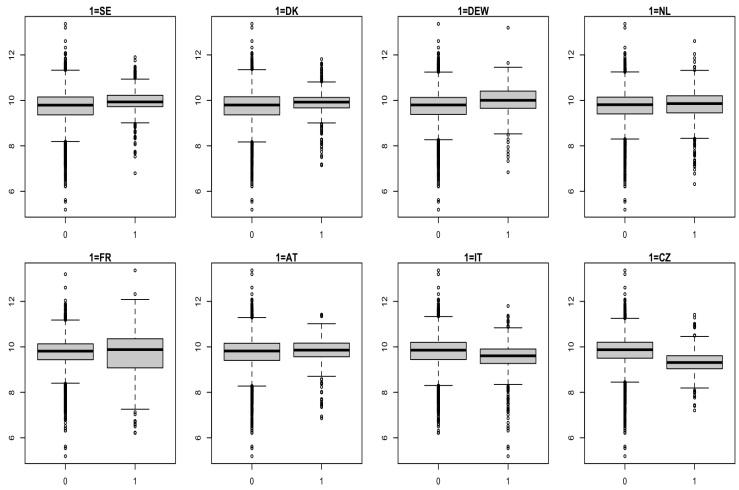
(b)

Figura 5.3: Grafici illustrativi della relazione marginale tra le variabili risposta (a) *Current wage* e (b) *Lifetime earnings* e, da sinistra verso destra: anni d'istruzione (*Years of edu*), coorte di nascita (*Cohort*) e provenienza da un'area rurale (*Rural*).

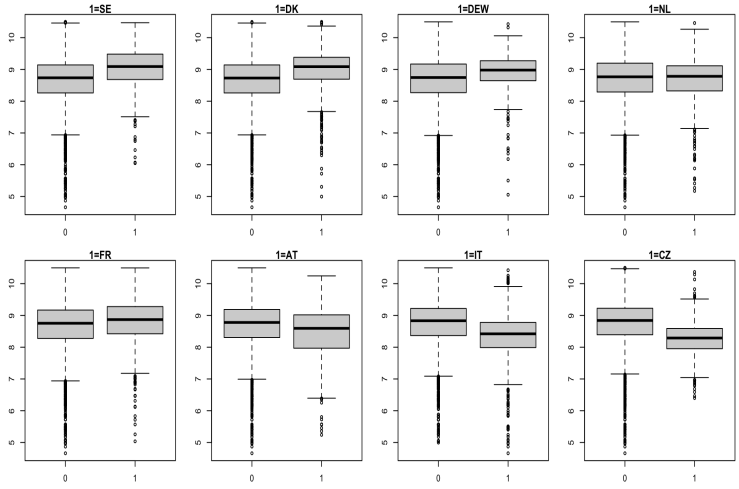
5.2.2 Analisi dei risultati

Brunello *et al.* (2017) conducono un'analisi su entrambe le variabili di reddito disponibili, per poi concentrarsi in un secondo momento sul reddito permanente *Lifetime earnings*, sul quale rilevano un effetto eterogeneo dell'istruzione al variare del background culturale della famiglia di origine (valutato dalla *proxy Book*). Invece, l'applicazione del modello (5.1) con risposta *Lifetime earnings*, non evidenzia differenze significative rispetto all'utilizzo di un più semplice OLS. Per questo, l'analisi che segue si concentra sull'*outcome Current wage* che ha prodotto risultati più interessanti.

La Tabella 5.1 riporta le stime dei parametri del modello lineare determinate con OLS, il modello a due stadi con l'inserimento di variabili strumentali di Brunello *et al.* (2017) e il miglior modello mistura selezionato. Quest'ultimo è stato ottenuto minimizzando il BIC al variare di diverse inizializzazioni dei



(a)



(b)

Figura 5.4: Distribuzioni empiriche delle variabili risposta (a) *Current wage* e (b) *Lifetime earnings* nei gruppi definiti dall'appartenenza ai diversi paesi europei, da sinistra verso destra: Svezia (SE), Danimarca (DK), Germania (DEW), Paesi Bassi (NL), Francia (FR), Austria (AT), Italia (IT), Repubblica Ceca (CZ).

	OLS	2SLS	Mistura		
			gruppo 1	gruppo 2	media 1 e 2
(Intercept)	9.286***	9.183***	5.390***	9.384***	9.137
Years of education	0.042***	0.052.	0.157***	0.045***	0.052
Rural	−0.041*	−0.032	0.291*	−0.049	−0.028
Cohort	0.004***	0.003	−0.000	−0.001	−0.001
SE	0.142***	0.153**	1.239**	0.141**	0.209
DK	0.019	0.018	1.265**	0.019	0.096
DEW	0.094*	0.085.	1.130**	0.092*	0.156
NL	−0.053	−0.045	0.886**	0.001	0.055
FR	−0.165***	−0.159***	1.284**	0.216***	0.282
AT	−0.041	−0.026	0.381	0.103*	0.120
IT	−0.183***	−0.150	1.398**	−0.115	−0.021
CZ	−0.549***	−0.553***	−0.458	−0.571**	−0.564

Tabella 5.1: Tabella coefficienti stimati con OLS, 2SLS e modello mistura a due gruppi, utilizzanto *Current wage* come variabile risposta.

parametri e al variare di G.

Modello lineare e modello a due stadi con variabili strumentali producono stime confrontabili, ma suggeriscono che ogni anno in più di istruzione porta rispettivamente ad un incremento medio di reddito del 4.2% e del 5.2%.

Relativamente alla metodologia proposta, il modello selezionato è una mistura a due componenti sbilanciate, con probabilità a priori π_g stimate pari a (0.075, 0.925). La stima dell'effetto relativo al gruppo predominante (gruppo 2) è molto vicina a quella ottenute mediante OLS, e tale coerenza è valida anche per la maggior parte dei parametri associati alle variabili di controllo. Il gruppo più piccolo (gruppo 1), invece, si caratterizza per un effetto del trattamento marcatamente superiore e stimato pari al 15.7%. Si noti come a livello di stima dell'effetto medio, l'ATE dato dai due gruppi del modello mistura (ottenuto come media ponderata degli effetti entro ciascun gruppo) corrisponde all'effetto di trattamento che Brunello *et al.* (2017) ottengono facendo ricorso alle variabili strumentali.

Per facilitare l'interpretazione dei gruppi, e la loro composizione, è stato stimato un albero di classificazione utilizzando l'etichetta di gruppo come variabile risposta e le variabili di controllo come esplicative. Il risultato è rappresentato in Figura 5.5. Si nota che le variabili che più discriminano i due gruppi sono il paese di appartenenza e la coorte di nascita dei soggetti. In particolare, appartengono al gruppo minore principalmente soggetti provenienti da Francia, Italia ed Austria e appartenenti alle generazioni meno recenti. I due gruppi invece si distinguono poco per composizione rispetto a percentuale di residenti in area rurale all'età di 10 anni (gruppo 1: 39%, gruppo 2: 45%).

Come discusso nel paragrafo 3.4, vale la pena sottolineare la dovuta cautela rispetto a questa interpretazione dei gruppi, per non trascurare gli eventuali

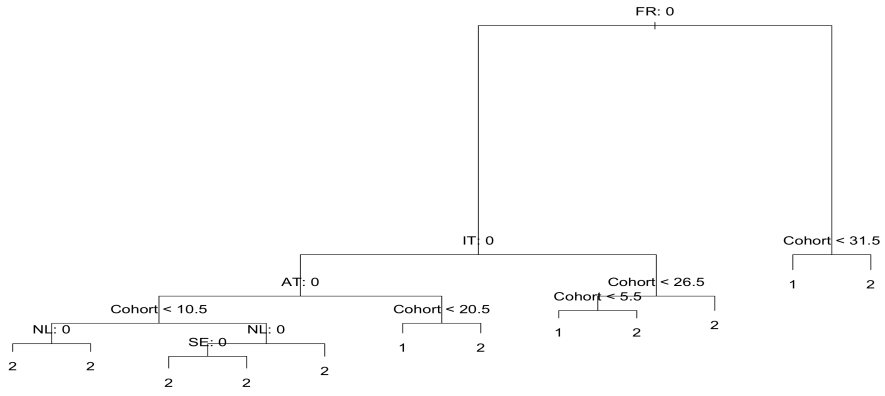


Figura 5.5: Albero di classificazione con etichetta di gruppo individuato dalla mistura come variabile risposta e variabili di controllo come esplicative.

rischi in cui si può incorrere nel caso di superficialità nella valutazione dei risultati.

Conclusioni

In questa tesi si è esplorata la possibilità di identificare eventuali fonti di eterogeneità latente mediante il ricorso a modelli di regressione a mistura finita per la valutazione di un effetto causale.

La metodologia proposta risulta essere uno strumento utile a questo scopo, semplice nella costruzione, poiché derivabile da una diretta estensione di metodi di stima e procedure inferenziali già esistenti nella letteratura statistica classica, e flessibile sia nella modellazione sia nella molteplicità di risultati che permette di ottenere.

Dalla valutazione della procedura, effettuata mediante un ampio studio di simulazione ed una applicazione a dati reali, è emerso che la metodologia risulta essere tanto più utile quanto più l'eterogeneità latente presente nella popolazione è associata ad una struttura a gruppi sufficientemente distinguibili. In caso contrario, i risultati che si ottengono non sono in generale peggiori di quelli che si otterrebbero senza il ricorso al suo utilizzo. Vale la pena notare, in proposito, che l'applicazione della metodologia proposta necessita, in linea di principio, della sola osservazione del trattamento e dell'*outcome* di interesse e dunque si configura come uno strumento interamente *data driven* e potenzialmente capace di ridurre la distorsione nella valutazione dell'effetto di un trattamento rafforzando l'efficacia di qualunque altro strumento proprio dell'inferenza causale.

D'altra parte, la ricerca di classi latenti nella distribuzione della risposta condizionata al trattamento (e ad eventuali variabili di controllo) non garantisce l'identificazione di un'eterogeneità pre-trattamento. Per questa ragione, i limiti della procedura si presentano nella misura in cui ai risultati che si ottengono non si accompagnano conoscenze teoriche del contesto e dei dati a disposizione, essenziali allo scopo di dare una corretta interpretazione delle stime e di mettere in guardia da potenziali errori di valutazione dei risultati. A questo proposito, l'applicazione di questa metodologia risulta essere tanto più utile quanto più i risultati ottenuti sono congruenti con un'ipotesi iniziale sostenuta da conoscenze pregresse del ricercatore.

È proprio in questa direzione che dovrebbe concentrarsi la ricerca futura, allo scopo di caratterizzare in modo più sostanziale in quali condizioni e sotto

quali assunzioni il modello individua eterogeneità corrispondente a un buon controllo, e non *bad control*.

Altri eventuali prossimi passi potrebbero essere di applicare la metodologia ad altre tipologie di dati come, ad esempio, dati panel o serie storiche, o di integrare il metodo proposto alle altre tecniche *data-driven* menzionate nei Capitoli precedenti.

Appendice A

Dettagli degli scenari di simulazione

Si riportano nel seguito i dettagli dello studio di simulazione illustrato nel Capitolo 4.

1. Scenario 1: caratteristica non osservata dicotomica $Z_i = \{Z_{i1}, Z_{i2}\}$, trattamento a due livelli $D_i = \{D_{i0}, D_{i1}\}$, effetto omogeneo $\beta_{1g} = \beta_1, g = 1, 2$:
   $\pi = (0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_1 = 200$.
   - assegnazione bilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{12} = 0.6$;
   - assegnazione sbilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{12} = 0.95$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = 55, \sigma_2^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = 70, \sigma_2^2 = 80$.
2. Scenario 2: caratteristica non osservata dicotomica $Z_i = \{Z_{i1}, Z_{i2}\}$, trattamento a due livelli $D_i = \{D_{i0}, D_{i1}\}$, effetto eterogeneo $\beta_{11} \neq \beta_{12}$:
   $\pi = (0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{11} = 200$, $\beta_{12} = -200$.
   - assegnazione bilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{12} = 0.6$;
   - assegnazione sbilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{12} = 0.95$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = 55, \sigma_2^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = 70, \sigma_2^2 = 80$.
3. Scenario 3: caratteristica non osservata a quattro livelli di eterogeneità, $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$, trattamento a due livelli $D_i = \{D_{i0}, D_{i1}\}$, effetto omogeneo $\beta_{1g} = \beta_1, g = 1, \ldots, 4$:

$\pi = (0.5, 0.5, 0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{03} = 1600$, $\beta_{04} = 1800$, $\beta_1 = 200$.

- assegnazione bilanciata dei trattamenti:
  $\rho_{11} = 0.3$, $\rho_{12} = 0.4$, $\rho_{13} = 0.6$, $\rho_{14} = 0.7$;
- assegnazione sbilanciata dei trattamenti:
  $\rho_{11} = 0.1$, $\rho_{12} = 0.4$, $\rho_{13} = 0.6$, $\rho_{14} = 0.95$;
- cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 60^2$;
- cluster eteroschedastici ben distinguibili: $\sigma_1^2 = \sigma_3^2 = 55, \sigma_2^2 = \sigma_4^2 = 65$;
- cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 75^2$;
- cluster eteroschedastici poco distinguibili: $\sigma_1^2 = \sigma_3^2 = 70, \sigma_2^2 = \sigma_4^2 = 80$.

4. Scenario 4: caratteristica non osservata a quattro livelli di eterogeneità, $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$, trattamento a due livelli $D_i = \{D_{i0}, D_{i1}\}$, effetto eterogeneo $\beta_{1g} \neq \beta_{1g'}, g \neq g'$:
   $\pi = (0.5, 0.5, 0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{03} = 1600$, $\beta_{04} = 1800$, $\beta_{1g} = 200$, $g = 1, 2, 3$, $\beta_{14} = -600$.
   - assegnazione bilanciata dei trattamenti:
     $\rho_{11} = 0.3$, $\rho_{12} = 0.4$, $\rho_{13} = 0.6$, $\rho_{14} = 0.7$;
   - assegnazione sbilanciata dei trattamenti:
     $\rho_{11} = 0.1$, $\rho_{12} = 0.4$, $\rho_{13} = 0.6$, $\rho_{14} = 0.95$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = \sigma_3^2 = 55, \sigma_2^2 = \sigma_4^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = \sigma_3^2 = 70, \sigma_2^2 = \sigma_4^2 = 80$.
5. Scenario 5: caratteristica non osservata dicotomica $Z_i = \{Z_{i1}, Z_{i2}\}$, trattamento a tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, effetto omogeneo $\beta_{jg} = \beta_j, g = 1, 2$, e lineare tra i livelli, $\beta_j = \beta_1$, $j = 1, 2$:
   $\pi = (0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_1 = 200$.
   - assegnazione bilanciata dei trattamenti:
     $\rho_{11} = 0.3$, $\rho_{21} = 0.2$, $\rho_{12} = 0.3$, $\rho_{22} = 0.5$;
   - assegnazione sbilanciata dei trattamenti:
     $\rho_{11} = 0.4$, $\rho_{21} = 0.1$, $\rho_{12} = 0.15$, $\rho_{22} = 0.8$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = 55, \sigma_2^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = 70, \sigma_2^2 = 80$.

6. Scenario 6: caratteristica non osservata dicotomica $Z_i = \{Z_{i1}, Z_{i2}\}$, trattamento a tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, effetto eterogeneo $\beta_{jg} \neq \beta_{jg'}, g \neq g'$, e lineare tra i livelli, $\beta_{jg} = \beta_{1g}$, $j = 1, 2$:
$\pi = (0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{11} = 200$, $\beta_{12} = -200$.
   - assegnazione bilanciata dei trattamenti:
     $\rho_{11} = 0.3$, $\rho_{21} = 0.2$, $\rho_{12} = 0.3$, $\rho_{22} = 0.5$;
   - assegnazione sbilanciata dei trattamenti:
     $\rho_{11} = 0.4$, $\rho_{21} = 0.1$, $\rho_{12} = 0.15$, $\rho_{22} = 0.8$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = 55, \sigma_2^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = 70, \sigma_2^2 = 80$.
7. Scenario 7: caratteristica non osservata a quattro livelli di eterogeneità, $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$, trattamento a tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, effetto omogeneo $\beta_{jg} = \beta_j, g = 1, \ldots, 4$, e lineare tra i livelli, $\beta_j = \beta_1$, $j = 1, 2$:
$\pi = (0.5, 0.5, 0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{03} = 1600$, $\beta_{04} = 1800$, $\beta_1 = 200$.
   - assegnazione bilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{21} = 0.2$, $\rho_{12} = 0.3$, $\rho_{22} = 0.4$, $\rho_{13} = 0.3$, $\rho_{23} = 0.5$, $\rho_{14} = 0.2$, $\rho_{24} = 0.7$;
   - assegnazione sbilanciata dei trattamenti: $\rho_{11} = 0.35$, $\rho_{21} = 0.05$, $\rho_{12} = 0.3$, $\rho_{22} = 0.4$, $\rho_{13} = 0.3$, $\rho_{23} = 0.5$, $\rho_{14} = 0.15$, $\rho_{24} = 0.8$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = \sigma_3^2 = 55, \sigma_2^2 = \sigma_4^2 = 65$;
   - cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 75^2$;
   - cluster eteroschedastici poco distinguibili: $\sigma_1^2 = \sigma_3^2 = 70, \sigma_2^2 = \sigma_4^2 = 80$.
8. Scenario 8: caratteristica non osservata a quattro livelli di eterogeneità, $Z_i = \{Z_{i1}, \ldots, Z_{i4}\}$, trattamento a tre livelli $D_i = \{D_{i0}, D_{i1}, D_{i2}\}$, effetto eterogeneo $\beta_{jg} \neq \beta_{jg'}, g \neq g'$, e lineare tra i livelli, $\beta_{jg} = \beta_{1g}$, $j = 1, 2$:
$\pi = (0.5, 0.5, 0.5, 0.5)$, con $\beta_{01} = 1200$, $\beta_{02} = 1400$, $\beta_{03} = 1600$, $\beta_{04} = 1800$, $\beta_{1g} = 200$, $g = 1, 2, 3$, $\beta_{14} = -200$.
   - assegnazione bilanciata dei trattamenti: $\rho_{11} = 0.3$, $\rho_{21} = 0.2$, $\rho_{12} = 0.3$, $\rho_{22} = 0.4$, $\rho_{13} = 0.3$, $\rho_{23} = 0.5$, $\rho_{14} = 0.2$, $\rho_{24} = 0.7$;
   - assegnazione sbilanciata dei trattamenti: $\rho_{11} = 0.35$, $\rho_{21} = 0.05$, $\rho_{12} = 0.3$, $\rho_{22} = 0.4$, $\rho_{13} = 0.3$, $\rho_{23} = 0.5$, $\rho_{14} = 0.15$, $\rho_{24} = 0.8$;
   - cluster omoschedastici ben distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 60^2$;
   - cluster eteroschedastici ben distinguibili: $\sigma_1^2 = \sigma_3^2 = 55, \sigma_2^2 = \sigma_4^2 = 65$;

- cluster omoschedastici poco distinguibili: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 75^2$;
- cluster eteroschedastici poco distinguibili: $\sigma_1^2 = \sigma_3^2 = 70, \sigma_2^2 = \sigma_4^2 = 80$.

Bibliografia

Angrist J. D.; Pischke J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Athey S.; Imbens G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113**(27), 7353–7360.

Azzalini A. (2001). *Inferenza statistica, una presentazione basata sul concetto di verosimiglianza*. Springer-Verlag Italia.

Bouveyron C.; Celeux G.; Murphy T.; Raftery A. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Brunello G.; Weber G.; Weiss C. T. (2017). Books are forever: Early life conditions, education and lifetime earnings in europe. *The Economic Journal*, **127**(600), 271–296.

Celeux G.; Fruewirth-Schnatter S.; Robert C. P. (2018). Model selection for mixture models - perspectives and strategies.

Chernozhukov V.; Demirer M.; Duflo E.; Fernández-Val I. (2023). Fischer-schultz lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india.

Cinelli C.; Forney A.; Pearl J. (2020). A crash course in good and bad controls. *ERN: Model Construction & Selection (Topic)*.

Dempster A. P.; Laird N. M.; Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Duflo E. (2018). Machinistas meet randomistas: useful ml tools for empirical researchers. *Presentation at the NBER Summer Institute*.

Faria S.; Soromenho G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, **80**(2), 201–225.

Fruhwirth-Schnatter S.; Celeux G.; Robert C. P. (2019). *Handbook of mixture analysis*. CRC press.

Gormley I.; Murphy T. (2011). Mixture of experts modelling with social science applications.

Hasan A.; Zhiyu W.; Mahani A. S. (2014). Fast estimation of multinomial logit models: R package mnlogit. *arXiv preprint arXiv:1404.3177*.

Imbens G. W. (2014). Instrumental Variables: An Econometrician's Perspective. *Statistical Science*, **29**(3), 323 – 358.

Kim K. (2020). Causal Inference with Complex Data Structures and Non-Standard Effects.

Kline R. B.; Little T. D. (2011). *Principles and practice of structural equation modeling / Rex B. Kline ; series editor's note by Todd D. Little*. Methodology in the social sciences. Guilford press, New York London, 3. ed. edizione.

McLachlan G.; Peel D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.

Pearl J. (2009). *Causality*. Cambridge University Press, Cambridge, UK, 2 edizione.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Titterington D. M.; Smith A. F. M.; Makov U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics. Wiley sons, Chichester.

Wager S.; Athey S. (2017). Estimation and inference of heterogeneous treatment effects using random forests.

Wilms R.; Mäthner E.; Winnen L.; Lanwehr R. (2021). Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, **5**, 100075.

Zorzetto D.; Bargagli-Stoffi F. J.; Canale A.; Dominici F. (2023). Confounder-dependent bayesian mixture model: Characterizing heterogeneity of causal effects in air pollution epidemiology.