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

Relazioni tra i parametri marginali e condizionati nei modelli di regressione logistica: una specificazione generale che coinvolge la distribuzione normale asimmetrica estesa

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche
Università degli Studi di Padova

Correlatore Prof. Elena Stanghellini
Dipartimento di Economia
Università degli Studi di Perugia

Laureando Matteo Gasparin
Matricola 2020317

Anno Accademico 2021/2022

Indice

Introduzione

In alcuni ambiti, come la sociologia, la psicologia o l'epidemiologia, risulta necessario comprendere la procedura mediante la quale una covariata va ad agire sulla variabile risposta. La *mediation analysis* è una metodologia utilizzata in questo contesto, con il fine di capire ed esplorare le relazioni tra la variabile risposta e le variabili indipendenti.

In questo lavoro, verranno introdotte le principali metodologie utilizzate nell'ambito dell'analisi di mediazione parametrica e le problematiche derivanti nel caso in cui la risposta abbia natura binaria. Successivamente verrà data una panoramica sulle distribuzioni asimmetriche, le quali verranno utilizzate nel definire i processi generatori utilizzati nel prosieguo. Nei Capitoli 3 e 4, si cercheranno di ricavare le relazioni che legano i parametri condizionati e marginali nei modelli di regressione logistica nel caso di covariate continue. Nel Capitolo 5 verranno applicate le metodologie introdotte ad una analisi di dati reali.

Capitolo 1

Mediation analysis

In statistica, un modello per l'analisi di mediazione cerca di identificare e spiegare il meccanismo o il processo alla base di una relazione osservata tra una variabile indipendente e una variabile dipendente attraverso l'inclusione di una o più variabili collegate sia con la risposta che con la covariata di interesse, e note con il nome di variabili mediatrici (MacKinnon, 2012). La *mediation analysis* è uno strumento importante nelle scienze comportamentali per studiare il ruolo delle variabili intermedie che risiedono nel percorso tra un trattamento e la variabile indicata come risposta. Questo tipo di analisi viene quindi utilizzata per valutare l'importanza relativa dei diversi percorsi attraverso i quali un trattamento può influenzare la risposta (VanderWeele, 2015). Questo filone di ricerca è molto comune negli studi sociali ed epidemiologici anche se trova applicazione in altri settori, come la psicologia e la salute pubblica. Nella sua formulazione più semplice, cioè nel caso di modelli per risposta continua con solamente 3 variabili, questa tematica risulta ben nota agli statistici in quanto presenta una stretta relazione con i coefficienti di regressione parziale e con il modello di regressione lineare multipla, come spiegato ad esempio in Iacobucci (2008). Tuttavia all'aumentare delle variabili e delle relazioni in gioco o al variare della natura delle variabili stesse, il problema si complica notevolmente. Nel prosieguo del Capitolo si farà una breve introduzione al fenomeno della mediazione e verranno esposte le principali metodologie utilizzate in ambito di regressione, sia per risposta continua che per risposta binaria.

1.1 Il fenomeno della *mediazione*

Uno dei metodi per spiegare delle relazioni causa-effetto è quello di capire come la causa agisce sulla risposta, ciò equivale a descrivere il meccanismo mediante il quale questa relazione si compie. In alcuni studi risulta infatti di interesse studiare come una causa e certe condizioni iniziali portino ad uno stato finale attraverso un processo comprendente diversi passi intermedi. Nella pratica può succedere infatti che una variabile intermedia sia responsabile di una parte degli effetti del trattamento sulla risposta; risulta quindi di interesse valutare quale porzione di effetti del trattamento sulla variabile risposta sia mediato tramite questa variabile. In particolare, gli effetti del trattamento che agiscono direttamente sulla risposta vengono identificati come *effetti diretti*, al contrario, quando gli effetti della covariata di interesse operano attraverso una variabile intermedia vengono definiti *effetti indiretti*. La somma tra le due tipologie di effetto, diretto ed indiretto, dà l'effetto totale del trattamento sulla variabile oggetto di studio.

Gran parte della letteratura si concentra sulla mediazione in ambito controfattuale (in relazione ai *potential outcome*). Ad esempio, siano X un trattamento binario e Y una risposta continua, allora le variabili Y_x sono definite controfattuali o risultati potenziali. Per ogni individuo, viene osservato solamente uno dei due risultati potenziali: Y_1 se l'individuo viene sottoposto al trattamento o Y_0 se l'individuo non viene sottoposto al trattamento. Se Y_1 risulta diverso da Y_0 allora si ha un effetto del trattamento sull'individuo pari a $Y_1 - Y_0$. In quanto per ogni persona non è possibile osservare entrambi i risultati potenziali non si può dare una stima dell'effetto a livello individuale. Si cerca quindi di dare una stima media per la popolazione, indicata tramite $\mathbb{E}[Y_1 - Y_0]$. Utilizzando modelli parametrici, risulta possibile esprimere tali effetti in termini di relazioni tra i parametri come descritto in VanderWeele (2015).

Per quanto l'analisi di mediazione aiuti a spiegare come un trattamento agisca su una variabile dipendente, al fine di trarre relazioni causa-effetto sono richieste assunzioni abbastanza stringenti, come spiegato ad esempio in VanderWeele (2016).

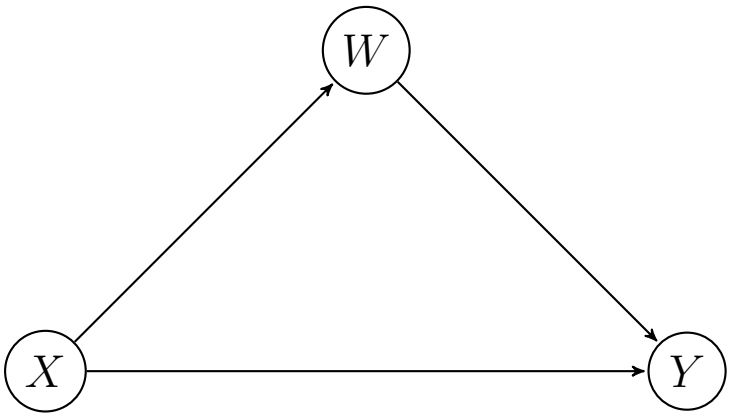

Figura 1.1: *Mediation analysis*: relazioni tra variabili

1.2 Approcci tradizionali

Come precedentemente anticipato, l'analisi di mediazione aiuta a comprendere il meccanismo mediante il quale una variabile indipendente va ad influire su una variabili dipendente. Le motivazioni principali che spingono un ricercatore ad applicare tale tipo di analisi sono principalmente di carattere interpretativo ed esplicativo, in quanto la comprensione del meccanismo può aiutare a confermare o a rifiutare una teoria, o semplicemente a migliorare la comprensione del fenomeno studiato. La letteratura relativa alla mediazione parametrica si è notevolmente ampliata nell'ultimo decennio, assumendo due approcci classici: il *metodo delle differenze* e il *metodo del prodotto*.

1.2.1 Il metodo delle differenze

Il metodo delle differenze consiste essenzialmente nello stimare due differenti modelli di regressione. Sia X la variabile indipendente di interesse, W un potenziale mediatore e Y la risposta. Il primo modello di regressione va a regredire semplicemente la variabile risposta Y sulla variabile X:

$$\mathbb{E}[Y \mid X = x] = \eta_0 + \eta_x x. \tag{1.1}$$

Il coefficiente η_x viene interpretato come effetto totale della variabile X sulla risposta Y. Il secondo modello di regressione va ad aggiungere all'equazione (1.1) il

mediatore, arrivando alla formulazione

$$\mathbb{E}[Y \mid X = x, W = w] = \beta_0 + \beta_x x + \beta_w w. \tag{1.2}$$

L'effetto indiretto o mediato, definito dalla sigla IE, è dato dalla differenza tra i due coefficienti di regressione relativi alla variabile X

$$IE = \eta_x - \beta_x.$$

Se il primo coefficiente di regressione, η_x, risulta considerevolmente più elevato in valore assoluto rispetto al coefficiente β_x, che corrisponde al coefficiente relativo alla variabile X nel modello con l'aggiunta di W, si ha una indicazione di mediazione in quanto una parte degli effetti del trattamento vengono mediati tramite l'aggiunta di una terza variabile.

L'effetto diretto, indicato dalla sigla DE, viene invece definito dal coefficiente di regressione parziale β_x,

$$DE = \beta_x,$$

poiché misura l'effetto sulla risposta rimanente anche con l'aggiunta del mediatore, e corrisponde al collegamento tra la variabile X e la variabile Y in Figura 1.1.

1.2.2 Il metodo del prodotto

Il metodo del prodotto viene descritto inizialmente in Cochran (1938) e viene successivamente ripreso dall'articolo di Baron & Kenny (1986), tale approccio viene per questo definito anche *metodo di Baron e Kenny* o *scomposizione di Cochran*. Anche questo metodo prevede l'utilizzo di due differenti modelli di regressione: il primo coincide con l'equazione (1.2), mentre il secondo utilizza come variabile dipendente il mediatore e come variabile esplicativa il trattamento

$$\mathbb{E}[W \mid X = x] = \theta_0 + \theta_x x. \tag{1.3}$$

Anche in questo caso l'effetto diretto è dato da β_x, mentre l'effetto indiretto è dato dal prodotto tra θ_x e β_w

$$IE = \theta_x \beta_w, \tag{1.4}$$

e può essere interpretato come l'effetto del trattamento sul mediatore moltiplicato per l'effetto del mediatore sulla risposta.

Ne deriva che l'effetto totale (TE) è dato dalla somma di due quantità che dipendono direttamente dai parametri delle equazioni (1.2) e (1.3)

$$TE = DE + IE = \beta_x + \theta_x \beta_w, \tag{1.5}$$

si noti che la nullità di alcuni dei parametri presenti in (1.2) e (1.3) porta all'annullamento di uno dei due addendi relativi all'effetto diretto o all'effetto indiretto. In particolare, la nullità di alcuni coefficienti coincide graficamente con la mancanza del collegamento tra le variabili interessate presenti in Figura (1.1).

Nel caso di risposta e mediatore continui con modelli di regressione lineare stimati ai minimi quadrati i due approcci coincidono, come mostrato in MacKinnon et al. (1995). Si può infatti scrivere

$$\begin{aligned} Y &= \beta_0 + \beta_x x + \beta_w w + \varepsilon_y \\ W &= \theta_0 + \theta_x x + \varepsilon_w \end{aligned}$$

con ε_y e ε_w errori gaussiani. Ne deriva

$$\begin{aligned} Y &= \beta_0 + \beta_x x + \beta_w(\theta_0 + \theta_x x + \varepsilon_w) + \varepsilon_y \\ &= \beta_0 + \theta_0 \beta_w + (\beta_x + \beta_w \theta_x) x + \varepsilon_y + \beta_w \varepsilon_w, \end{aligned}$$

che dimostra $\eta_x = \beta_x + \theta_x \beta_w$, cioè che l'effetto indiretto calcolato tramite i due metodi coincide in quanto $\eta_x - \beta_x = \theta_x \beta_w$.

1.2.3 Permettere l'interazione tra regressori

In alcuni casi è inoltre presente una interazione tra la variabile X e il mediatore W, e risulta necessario andare a scomporre l'effetto totale in effetti diretti ed effetti indiretti. Si supponga quindi che nel modello per la risposta sia presente un termine di interazione tra il mediatore e il trattamento

$$\mathbb{E}[Y \mid X = x, W = w] = \beta_0 + \beta_x x + \beta_w w + \beta_{xw} xw, \tag{1.6}$$

mentre il modello per il mediatore rimanga invariato e pari a

$$\mathbb{E}[W \mid X = x] = \theta_0 + \theta_x x.$$

Nel caso di corretta specificazione del modello le stime degli effetti diretti ed indiretti per un cambio del trattamento da x a x' sono date da

$$\begin{aligned} DE &= \{\beta_x + \beta_{xw}(\theta_0 + \theta_x x')\}(x - x'), \\ IE &= \theta_x(\beta_w + \beta_{xw})(x - x'). \end{aligned} \tag{1.7}$$

L'effetto totale è dato dalla somma delle due componenti, molto spesso viene inoltre utilizzata come misura la proporzione di effetto mediato, ottenuta tramite il rapporto tra l'effetto indiretto e l'effetto totale, come descritto in VanderWeele (2016). Si noti che se il parametro β_{xw} risulta pari a 0, allora si torna alla casistica descritta nel Paragrafo 1.2.2.

1.3 Mediazione nel caso di dati binari

Le precedenti formulazioni sono ristrette al caso dei modelli di regressione lineare, infatti il metodo del prodotto e il metodo delle differenze portano agli stessi risultati solamente nel caso in cui sia la risposta che il mediatore presentano natura continua. Tuttavia alcune variabili spiegate non possono essere trattate tramite modelli di regressione lineare in quanto di natura discreta o qualitativa, risulta quindi necessario ampliare le metodologie "tradizionali" per il calcolo dell'effetto totale (metodo del prodotto e metodo delle differenze) ai contesti non lineari. Si cercano quindi di sviluppare dei metodi facilmente interpretabili che consentano di suddividere in maniera semplice l'effetto totale in effetto diretto ed indiretto e che presentino proprietà simili alla formula presentata in equazione (1.5) e alla sua generalizzazione derivata in equazione (1.7). Particolare attenzione viene riposta nel caso in cui la risposta abbia natura binaria: tale casistica si riscontra non solo quando la variabile sia di per sé dicotomica (ad esempio *si/no*), ma anche nei casi in cui la risposta viene ottenuta tramite dicotomizzazione di una sottostante variabile latente quantitativa.

Nel caso di dati binari risulta quindi necessario andare a generalizzare le equa-

zioni definite nel Paragrafo 1.2.1, in particolare l'equazione (1.2) diviene

$$g\{\mathbb{E}[Y \mid X = x, W = w]\} = \beta_0 + \beta_x x + \beta_w w, \tag{1.8}$$

mentre per la (1.1) si ha

$$g\{\mathbb{E}[Y \mid X = x]\} = \eta_0 + \eta_x x, \tag{1.9}$$

dove $g(\cdot)$ è una funzione con codominio reale, monotona crescente e viene definita con il nome di funzione legame o *link function*. Come descritto in Lin et al. (1998) ed in Greenland et al. (1999), una delle principali difficoltà nel caso di dati con risposta dicotomica, risiede nel fatto che se il modello definito in (1.8) risulta correttamente specificato, non è detto che il modello marginale presentato in (1.9) sia lineare rispetto alla variabile x. Ad esempio, assumendo per l'equazione (1.8) una funzione legame logistica ed un mediatore continuo, si ottiene che la relazione marginale tra Y ed X è data da

$$\mathbb{P}(Y = 1 \mid X = x) = \int_{-\infty}^{+\infty} \frac{\exp(\beta_0 + \beta_x x + \beta_w w)}{1 + \exp(\beta_0 + \beta_x x + \beta_w w)} f(w \mid X = x)\, dw,$$

dove l'integrale risulta valutabile esplicitamente solamente in casi notevoli (MacKinnon et al., 2007). Gran parte delle letteratura introduce tuttavia una regressione logistica standard anche per il modello modello marginale definito in (1.9), nel prosieguo si cercherà quindi di valutare quando questa assunzione può essere adeguata.

1.3.1 Un'approssimazione per risposta binaria e mediatore continuo

Un importante risultato utilizzato nel contesto dei dati binari si basa sull'articolo di VanderWeele & Vansteelandt (2010), in particolare sia W un mediatore normalmente distribuito, Y una risposta binaria e sia la classe $Y = 1$ relativamente rara (solitamente si sceglie il 10% come limite). Si definisca il seguente modello di regressione logistica per la risposta

$$\log \frac{\mathbb{P}(Y = 1 \mid X = x, W = w)}{\mathbb{P}(Y = 0 \mid X = x, W = w)} = \beta_0 + \beta_x x + \beta_w w,$$

ed un modello di regressione lineare per il mediatore

$$\mathbb{E}[W \mid X = x] = \theta_0 + \theta_x x.$$

Se il modello risulta correttamente specificato allora risulta possibile ottenere delle approssimazioni per effetti diretti ed indiretti

$$\begin{aligned} DE &\approx \exp\{\beta_x\}, \\ IE &\approx \exp\{\beta_w \theta_x\}, \end{aligned}$$

dove queste approssimazioni sono espresse in termini di log-rapporto di quote in quanto per l'equazione (1.8) si è scelta la funzione legame logistica. Si noti che gli effetti diretti risultano pari ad $\exp(\beta_x)$ mentre gli effetti indiretti sono $\exp(\theta_x \beta_w)$; si ottengono quindi delle espressioni molto simili a quelle derivanti dal metodo del prodotto. Si dimostra inoltre che solo nel caso in cui la risposta sia rara il metodo del prodotto ed il metodo delle differenze coincidono approssimativamente in quanto il corrispettivo dell'equazione (1.9) risulta

$$\begin{aligned} \mathbb{E}[Y \mid X = x] &= \mathbb{P}(Y = 1 \mid X = x) \\ &\approx \exp(\eta_0 + \eta_x x), \end{aligned} \tag{1.10}$$

tuttavia se la risposta non è rara i due metodi divergono. Un'alternativa da utilizzare per aggirare questo problema è utilizzare un modello log-binomiale, assumendo $g(\cdot) = \log(\cdot)$ nelle equazioni (1.8) ed (1.9). Utilizzando questo approccio gli effetti diretti ed indiretti saranno espressi in termini di rischio relativo.

1.4 Le problematiche relativi ai modelli logistici annidati

In molti studi sociologici ed epidemiologici la variabile risposta Y viene definita tramite discretizzazione di una variabile latente continua Y^*. Date due covariate continue X e W, risulta logico andare a postulare una formulazione lineare per il logit della risposta, che può essere ottenuta a partire dal seguente modello per la variabile latente,

$$Y^* = \beta_x^* x + \beta_w^* w + e_F \tag{1.11}$$

con $e_F = \sigma_F \varepsilon_F$, dove ε_F si distribuisce come una variabile aleatoria logistica standard definita in Appendice A.3. Nel caso di dati binari, viene osservata solamente una discretizzazione di Y^*, in particolare

$$Y = \begin{cases} 1 & \text{se } Y^* > \tau, \\ 0 & \text{altrimenti,} \end{cases} \tag{1.12}$$

dove τ solitamente risulta pari a 0. Assumendo senza perdita di generalità $\tau = 0$, si ha che il modello con ambedue i regressori, denominato *full model* risulta

$$\begin{aligned} \mathbb{P}(Y = 1 \mid X = x, W = w) &= \mathbb{P}(Y^* > 0 \mid X = x, W = W) \\ &= \mathbb{P}\left(\varepsilon_F > \frac{-\beta_x^* x - \beta_w^* w}{\sigma_F}\right) \\ &= \frac{\exp(\beta_x x + \beta_w w)}{1 + \exp(\beta_x x + \beta_w w)}, \end{aligned} \tag{1.13}$$

questa rappresentazione giustifica l'utilizzo di un predittore lineare per il logit di Y, dove i coefficienti β_x e β_w risultano rispettivamente

$$\beta_x = \frac{\beta_x^*}{\sigma_F}, \quad \beta_w = \frac{\beta_w^*}{\sigma_F}. \tag{1.14}$$

Nell'articolo di Karlson et al. (2012) viene studiato il caso in cui anche il modello ridotto, derivante dall'omissione della variabile W tra i regressori, risulta lineare con errore avente distribuzione logistica rappresentato come segue

$$Y^* = \eta_x^* x + u_R \tag{1.15}$$

con $u_R = \sigma_R \nu_R$, dove ν_R possiede distribuzione logistica standard. In questo caso il coefficiente relativo alla variabile X risulta pari a

$$\eta_x = \frac{\eta_x^*}{\sigma_R},$$

e risulta evidente che il metodo delle differenze non risulta utilizzabile in quanto le diversità non riguardano solamente gli effetti dovuti all'inclusione o meno del mediatore, ma sono presenti anche delle differenze dovute ai differenti parametri

di scala poiché

$$\beta_x - \eta_x = \frac{\beta_x^*}{\sigma_F} - \frac{\eta_x^*}{\sigma_R} \neq \beta_x^* - \eta_x^*.$$

Risulta tuttavia necessario precisare che la distribuzione dell'errore u_R dipende, oltre che dalla distribuzione dell'errore e_F presente in equazione (1.11), anche dalla distribuzione del mediatore W. Ne deriva che i due modelli non differiscono solo a causa della diversa scala ma anche dalle distribuzioni dei loro termini di errore. Quanto appena spiegato può avere conseguenze nel confronto tra modelli annidati in quanto la forma funzionale definita in (1.15) può risultare poco appropriata per il modello. Nel caso di mediatore continuo, MacKinnon et al. (2007) suggerisce quindi l'utilizzo del metodo basato sul prodotto di coefficienti nel caso della regressione logistica, in quanto porta asintoticamente a dei risultati migliori rispetto al metodo delle differenze. Il metodo del prodotto presenta inoltre la proprietà di essere facilmente calcolabile, e permette di dividere facilmente effetti diretti ed indiretti.

1.5 Possibili sviluppi

Come già sottolineato in precedenza, studiare gli effetti dovuti all'aggiunta o alla rimozione di una variabile può risultare utile al fine di quantificare l'effetto dovuto a variabili intermedie e che può essere rimosso dopo essersi condizionati ai loro valori. Nei prossimi Capitoli si cercherà di approfondire il tema della *mediation analysis* applicata al caso di dati con risposta binaria, in particolare, risulterà d'interesse andare a studiare la relazione tra parametri condizionati e marginali utilizzando dei meccanismi generatori di dati specifici. Lo scopo sarà quello di ottenere delle relazione interpretabili che presentino la proprietà di scomporre l'effetto totale in termini di effetti diretti ed indiretti.

Capitolo 2

Le distribuzioni *skew-symmetric*

Questo Capitolo servirà a presentare alcune distribuzioni di densità asimmetriche che saranno utili nel definire specifici processi generatori dei dati a cui verranno applicati i metodi presentati nel Capitolo 1. Lo studio delle distribuzioni di probabilità parametriche nasce dall'esigenza di descrivere la realtà di interesse attraverso delle espressioni che siano matematicamente trattabili e che attraverso un numero finito di parametri siano in grado di rappresentare il meccanismo generatore dei dati sottostante al fenomeno oggetto di studio. In particolare, la distribuzione di probabilità deve considerarsi una verosimile astrazione della realtà e deve presentare una buona flessibilità ai dati.

In generale, l'introduzione di varianti asimmetriche per le usuali distribuzioni, come la normale o la t di Student, nasce con il fine di ricercare metodi flessibili volti a ridurre alcuni elementi di rigidità delle distribuzioni di partenza. In alcune scienze, come la finanza o la biologia, risulta utile andare ad aggiungere alle usuali densità delle varianti che permettano una maggiore adattabilità ai dati, in quanto eventi positivi e negativi non si verificano in egual modo. La costruzione di un insieme di distribuzioni asimmetriche spesso deriva dalla perturbazione di densità simmetriche e tale insieme contiene al proprio interno, come caso particolare, la versione simmetrica della distribuzione. A partire dal lavoro di de Helguero (1908), in letteratura sono stati presentati vari studi su queste famiglie; in questo Capitolo si farà riferimento al filone sviluppato in seguito all'articolo di Azzalini (1985).

2.1 Costruzione di una densità asimmetrica

Come primo passo risulta necessario andare a definire il concetto di simmetria, tale nozione infatti gioca un ruolo centrale nello sviluppo del Capitolo. Nel caso univariato si dice che una densità f_0 è simmetrica rispetto al valore x_0 se $f_0(x - x_0) = f_0(x_0 - x)$ per tutti gli x; solitamente si assume senza perdita di generalità $x_0 = 0$. Nel caso d-dimensionale, esistono varie definizioni di densità simmetrica: la più utilizzata richiede che la variabile casuale X sia distribuita come la variabile casuale $-X$. Nel caso in cui X sia continua con densità $f_0(x)$ tale condizione richiede, similmente al caso unidimensionale, che $f_0(x) = f_0(-x)$ per ogni $x \in \mathbb{R}^d$ (Serfling, 2004).

Azzalini (1985) introduce uno schema generale che, a partire da una funzione di densità simmetrica chiamata *densità di base*, permette di definire un insieme di densità semplicemente perturbando attraverso un fattore tale densità di base.

Lemma 1 *(Azzalini, 1985) Siano f_0 una funzione di densità di probabilità in $\mathbb{R}^d$, $G_0(\cdot)$ una funzione di ripartizione definita sull'asse reale e $w(\cdot)$ una funzione $\mathbb{R}^d \to \mathbb{R}$, tali che*

$$f_0(x) = f_0(-x), \quad w(-x) = -w(x), \quad G_0(-z) = 1 - G_0(z)$$

per ogni $x \in \mathbb{R}^d$, $z \in \mathbb{R}$. Allora,

$$f(x) = 2f_0(x)G_0\{w(x)\} \tag{2.1}$$

è una funzione di densità in $\mathbb{R}^d$.

Tale procedura è generale e permette la costruzione di densità univariate e multivariate semplicemente modificando una funzione di densità di base attraverso un fattore che può essere scelto abbastanza liberamente in quanto deve rispettare proprietà molto semplici. La famiglia definita in equazione (2.1) gode inoltre della proprietà di riflessione: infatti se X ha distribuzione (2.1), allora $-X$ possiede lo stesso tipo di distribuzione con $w(x)$ sostituito da $-w(x)$.

Il fattore $G_0\{w(x)\}$ può modificare profondamente e in modi molto diversi tra loro la densità di base, per il prosieguo si farà spesso riferimento al caso più specifico

in cui $w(x) = \alpha^\top x$ con $\alpha \in \mathbb{R}^d$, in quanto il più comune ed utile nella pratica. Nel caso unidimensionale la costante α regola l'asimmetria della distribuzione, da qui il nome *skew-symmetric*. Si noti inoltre che nel caso in cui il parametro α sia nullo allora $f(x) = f_0(x)$.

2.2 Rappresentazione stocastica

Una delle proprietà più interessanti delle famiglie *skew-symmetric* è che esse ammettono una varietà di rappresentazioni stocastiche, che ne giustificano l'utilizzo come modello nel caso di dati osservati.

Lemma 2 *(Azzalini & Capitanio, 2013) Sia Z_0 una variabile aleatoria d-dimensionale con densità $f_0(x)$ e sia $T \sim G_0$, con T indipendente da Z_0. Si definisca quindi,*

$$S = \begin{cases} 1 & \text{se } \; w(Z_0) - T > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

allora la variabile $Z = (Z_0 \mid S = 1)$ ha densità definita in equazione (2.1).

Si noti che la densità della variabile casuale $w(Z_0) - T$ risulta essere nota solamente in casi notevoli, tuttavia risulta semplice calcolarne i momenti essendo somma di variabili aleatorie indipendenti.

Tale rappresentazione stocastica presenta dei risvolti pratici anche nel caso in cui si voglia simulare da una variabile casuale asimmetrica così definita. Risulta sufficiente infatti simulare da Z_0 e T indipendentemente e successivamente mantenere solamente i valori tali per cui viene rispettata la condizione $w(Z_0) - T > 0$.

2.3 La distribuzione normale asimmetrica univariata

Tra le varie famiglie di distribuzioni che si possono generare dalla densità definita in equazione (2.1) sicuramente la più studiata ed utilizzata risulta essere la distribuzione normale asimmetrica, definita anche *skew-normal*.

In particolare tale distribuzione si ottiene scegliendo $\varphi(\cdot)$ come densità di base e $G_0(\cdot) = \Phi(\cdot)$, con $\varphi(\cdot)$ e $\Phi(\cdot)$ funzione di densità e di ripartizione di una normale

standardizzata, la densità di probabilità così ottenuta risulta

$$f(z;\alpha) = 2\varphi(z)\Phi(\alpha z) \quad -\infty < z < \infty, \tag{2.2}$$

dove α appartiene ai numeri reali. Tale distribuzione viene definita normale asimmetrica standard di parametro α ed indicata con $Z \sim SN(0,1,\alpha)$. L'espressione definita in (2.2) viene solitamente generalizzata andando ad introdurre i parametri di posizione e scala. Sia infatti $Y = \xi + \omega Z$ dove $\xi \in \mathbb{R}$ e $\omega \in \mathbb{R}^+$ allora la variabile casuale Y ha densità

$$f(y;\xi,\omega,\alpha) = \frac{2}{\omega}\varphi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\alpha\frac{y-\xi}{\omega}\right) \quad -\infty < y < \infty, \tag{2.3}$$

e verrà indicata tramite $Y \sim SN(\xi,\omega^2,\alpha)$.

Ci sono alcune semplici proprietà che derivano dalla definizione e dalle caratteristiche descritte nella Sezione 2.1. In particolare, se $Z \sim SN(0,1,\alpha)$ con densità $f(z;\alpha)$ allora:

1. per $\alpha = 0$ ci si riconduce ad una distribuzione normale standardizzata, infatti $f(z;0) = \varphi(z)$ per ogni z;

2. $f(0;\alpha) = \varphi(0)$ per ogni α;

3. $-Z \sim SN(0,1,-\alpha)$ per ogni z;

4. per $\alpha \to \infty$ ci si riconduce alla distribuzione *half-normal* di densità $2\varphi(z)$, con $z \geq 0$.

Si noti in Figura 2.1 come cambia la curva definita in (2.2) al variare del parametro α. Per valori negativi di α, a seguito della proprietà 3 avremo una distribuzione speculare rispetto all'asse verticale.

Facendo riferimento alla rappresentazione stocastica descritta nel Paragrafo 2.2, la variabile $Z \sim SN(0,1,\alpha)$ può essere ottenuta attraverso la rappresentazione

$$Z = (X_0 \mid \alpha X_0 - T > 0), \tag{2.4}$$

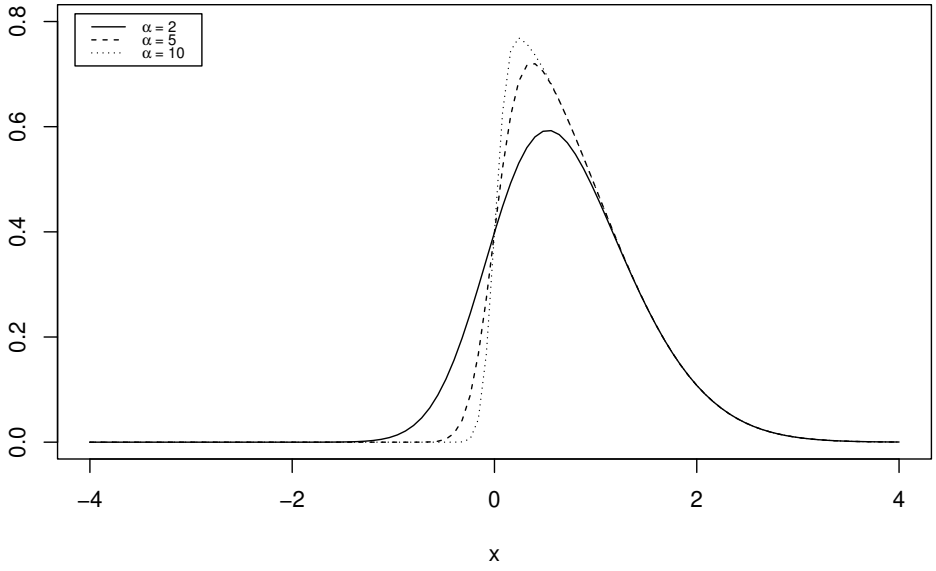

Figura 2.1: Grafico di una distribuzione $SN(0, 1, \alpha)$ per alcuni valori di α.

dove X_0 e T sono $N(0, 1)$ indipendenti. Risulta possibile esprimere tale costruzione a partire da una normale bivariata (X_0, X_1) con marginali standard dove

$$X_1 = \frac{\alpha X_0 - T}{\sqrt{1 + \alpha^2}}$$

tale che $\text{corr}(\text{X}_0, \text{X}_1) = \delta(\alpha) = \frac{\alpha}{\sqrt{1+\alpha^2}}$. La rappresentazione stocastica definita in (2.4) risulta equivalente a

$$Z = (X_0 \mid X_1 > 0).$$

Questa seconda rappresentazione, nonostante sia equivalente alla prima, risulta più attraente da un punto di vista modellistico infatti in alcuni casi si ha che una certa variabile $X_0^{'}$ risulta osservata solamente se un'altra variabile $X_1^{'}$, correlata con la prima, eccede una certa soglia; tale situazione viene definita come *campione selettivo*. In particolare se $(X_0^{'}, X_1^{'})$ sono congiuntamente normali con varianza unitaria e la soglia corrisponde al valor medio di $X_1^{'}$ allora siamo di fronte al caso sopracitato. Successivamente verrà trattato il caso in cui la soglia può assumere un valore arbitrario nel dominio della funzione. Altri metodi per rappresentare e

per generare da una distribuzione normale asimmetrica sono descritti in Azzalini & Capitanio (2013).

2.3.1 Momenti

Per il calcolo dei momenti risulta utile richiamare il risultato definito in Zacks (1981).

Lemma 3 *(Zacks, 1981) Siano* $U \sim N(0,1)$ *e* $h, k \in \mathbb{R}$ *allora*

$$\mathbb{E}[\Phi(hU+k)] = \Phi\left(\frac{k}{\sqrt{1+h^2}}\right)$$

A questo punto risulta semplice andare a calcolare la funzione generatrice dei momenti di $Y \sim SN(\xi, \omega^2, \alpha)$ che risulta

$$\begin{aligned} M_y(t) &= \mathbb{E}[\exp(\xi t + \omega Z t)] \\ &= 2\exp\left(\xi t + \frac{1}{2}\omega^2 t^2\right)\Phi(\delta\omega t), \end{aligned}$$

con $\delta \in (-1,1)$ precedentemente definito. Nel caso standard, con $\xi = 0$ e $\omega = 1$ si ha che i primi quattro momenti risultano

$$\begin{aligned} \mathbb{E}[Y] &= \sqrt{\frac{2}{\pi}}\delta, \\ \mathbb{V}ar[Y] &= 1 - \frac{2}{\pi}\delta^2, \\ \gamma_1[Y] &= \frac{4-\pi}{2}\text{sign}(\alpha)\left[\frac{\mathbb{E}^2[Y]}{\mathbb{V}ar[Y]}\right]^{\frac{3}{2}}, \\ \gamma_2[Y] &= 2(\pi-3)\left[\frac{\mathbb{E}^2[Y]}{\mathbb{V}ar[Y]}\right]^2, \end{aligned}$$

dove γ_1 e γ_2 rappresentano rispettivamente l'indice di asimmetria e l'indice di curtosi.

2.4 La distribuzione normale asimmetrica multivariata

Una naturale estensione della normale asimmetrica al caso d-dimensionale è stata introdotta in Azzalini & Dalla Valle (1996), mentre la sua applicazione è successivamente stata studiata nel lavoro di Azzalini & Capitanio (1999). Una variabile casuale d-dimensionale Z ha distribuzione normale asimmetrica d-dimensionale, indicata con $Z \sim SN_d(0, \bar{\Omega}, \alpha)$, se ha funzione di densità del tipo

$$f(z; \bar{\Omega}, \alpha) = 2\varphi_d(z; \bar{\Omega})\Phi(\alpha^\top z), \quad z \in \mathbb{R}^d \tag{2.5}$$

dove α è un vettore d-dimensionale, $\bar{\Omega}$ è una matrice di correlazione di dimensioni $d \times d$ e $\varphi_d(\cdot; A)$ rappresenta la densità di una normale multivariata con matrice di varianze e covarianze pari ad A. Anche in questo caso risulta possibile estendere la funzione di densità in (2.5) aggiungendo i parametri di posizione e scala. Si definiscano quindi $\xi \in \mathbb{R}^d$ e $\omega = \text{diag}(\omega_{z_1}, \ldots, \omega_{z_d})$ matrice diagonale con elementi positivi, ne deriva che la variabile $Y = \xi + \omega Z \sim SN(\xi, \Omega, \alpha)$ ha densità di probabilità

$$f(y; \xi, \Omega, \alpha) = 2\varphi_d(y - \xi; \Omega)\Phi(\alpha^\top \omega^{-1}(y - \xi)), \quad y \in \mathbb{R}^d,$$

con $\Omega = \omega\bar{\Omega}\omega$. Risulta utile notare che usando questa notazione si assume implicitamente che Ω sia definita positiva, vale inoltre la relazione $\omega = (\Omega \odot I_d)^{1/2}$ dove $\odot$ indica il prodotto di Hadamard elemento per elemento definito in Appendice A.1.

2.4.1 Rappresentazione stocastica e momenti

Similmente a quanto descritto nel caso univariato, la rappresentazione stocastica della variabile *skew-normal* multidimensionale può essere ottenuta a partire da una distribuzione normale multivariata. Come spiegato in Capitanio et al. (2003), sia

$$\begin{pmatrix} X_0 \\ X_1 \end{pmatrix} \sim N_{d+1}(0, \bar{\Omega}^*), \quad \text{con } \bar{\Omega}^* = \begin{pmatrix} \bar{\Omega} & \delta \\ \delta^\top & 1 \end{pmatrix} \tag{2.6}$$

dove δ è un vettore contenente la correlazione tra X_0 e X_1 mentre $\bar{\Omega}^*$ è una matrice di correlazione a rango pieno. Ne deriva che la relazione tra α e δ risulta biunivoca,

in particolare si ottiene

$$\alpha = (1 - \delta^\top \bar{\Omega}^{-1}\delta)^{-1/2}\bar{\Omega}^{-1}\delta, \tag{2.7}$$

$$\delta = (1 + \alpha^\top \bar{\Omega}\alpha)^{-1/2}\bar{\Omega}\alpha. \tag{2.8}$$

La variabile $Z \sim SN_d(0, \bar{\Omega}, \alpha)$ può essere rappresentata come

$$Z = (X_0 \mid X_1 > 0), \tag{2.9}$$

questa rappresentazione mette in risalto la relazione tra la distribuzione normale asimmetrica e il sottostante meccanismo di censura di una variabile normale multivariata, abbastanza comune nei contesti applicativi soprattutto nelle scienze sociali. Tale procedura, come precedentemente anticipato, prende infatti il nome di campione selettivo in quanto una variabile viene osservata solamente se un'altra variabile ad essa correlata rispetta una determinata condizione. Tale procedimento verrà esteso nei Capitoli successivi ed applicato al caso della discretizzazione di variabili latenti continue.

Per il calcolo della funzione generatrice dei momenti è necessario andare ad estendere il Lemma 3 al caso multivariato.

Lemma 4 *Siano $U \sim N_d(0, \Sigma)$, k un numero reale e $h \in \mathbb{R}^d$ allora*

$$\mathbb{E}[\Phi(h^\top U + k)] = \Phi\left(\frac{k}{\sqrt{1 + h^\top \Sigma h}}\right) \tag{2.10}$$

Utilizzando questo risultato congiuntamente alla regola del completamento del quadrato si ottiene che la funzione generatrice dei momenti di $Y = \xi + \omega Z \sim SN_d(\xi, \Omega, \alpha)$ risulta

$$M_y(t) = 2\exp\left(t^\top \xi + \frac{1}{2}t^\top \Omega t\right)\Phi(\delta^\top \omega t), \quad t \in \mathbb{R}^d. \tag{2.11}$$

2.4.2 Distribuzione marginale

Dal calcolo della funzione generatrice dei momenti segue che la distribuzione *skew-normal* multidimensionale è chiusa rispetto alla marginalizzazione. Sia $Z \sim SN_d(\xi, \Omega, \alpha)$, partizionata in $Z^\top = (X^\top, W^\top)$, di dimensioni h e $d - h$

rispettivamente con

$$\xi = \begin{pmatrix} \xi_x \\ \xi_w \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{xx} & \Omega_{xw} \\ \Omega_{wx} & \Omega_{ww} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_x \\ \alpha_w \end{pmatrix}, \quad \delta = \begin{pmatrix} \delta_x \\ \delta_w \end{pmatrix}, \tag{2.12}$$

corrispettivi partizionamenti di ξ, Ω, α e δ. Valutando la (2.11) in $t^\top = (s^\top, 0)$ si ottiene che la funzione che la funzione generatrice dei momenti di X è

$$M_x(s) = 2 \exp\left(s^\top \xi_x + \frac{1}{2} s^\top \Omega_{xx} s\right) \Phi(\delta_x^\top \omega_{xx} s), \quad s \in \mathbb{R}^h. \tag{2.13}$$

che risulta essere pari alla funzione generatrice dei momenti di una SN con parametro di posizione ξ_x e parametro di scala Ω_{xx}, dove $\omega_{xx} = (\Omega_{xx} \odot I_h)^{1/2}$. Dopo alcuni passaggi algebrici si ottiene, come dimostrato in Azzalini & Dalla Valle (1996), che

$$X \sim SN_h(\xi_x, \Omega_{xx}, \alpha_{x(w)}), \tag{2.14}$$

con

$$\alpha_{x(w)} = (\alpha_x + \bar{\Omega}_{xx}^{-1} \bar{\Omega}_{xw} \alpha_w)(1 + \alpha_w^\top \bar{\Omega}_{ww.x} \alpha_w)^{-1/2}, \tag{2.15}$$

e $\bar{\Omega}_{ww.x} = \bar{\Omega}_{ww} - \bar{\Omega}_{wx} \bar{\Omega}_{xx}^{-1} \bar{\Omega}_{xw}$. Si nota quindi che il j-esimo elemento di α della distribuzione congiunta non coincide con il medesimo elemento della distribuzione marginale. Al contrario dalla (2.13) si nota che il parametro δ_x risulta lo stesso sia per la distribuzione marginale che per la distribuzione congiunta.

2.5 Estensione delle famiglie asimmetriche

Risulta possibile ed utile estendere la famiglia di distribuzioni definita in (2.1) al fine di rilassare alcuni vincoli ed arrivare ad una famiglia di densità più generale. Nonostante la classe definita in (2.1) risulti molto ampia esistono alcune ragioni per derivare una formulazione più generale: le principali motivazioni sono indotte con il fine di ottenere un maggiore adattamento ai dati osservati. Risulta per questo utile aggiungere alla funzione $w(x)$ un parametro reale α_0 all'interno della funzione di perturbazione $G_0(\cdot)$. Come mostrato in Azzalini & Capitanio (2013), segue che

$$f(x) = f_0(x) \frac{G_0\{\alpha_0 + w(x)\}}{\mathbb{P}\{w(Z_0) - T > -\alpha_0\}} \tag{2.16}$$

è una densità in $\mathbb{R}^d$. Tale distribuzione viene definita *asimmetrica estesa*, in quanto estensione del caso con $\alpha_0 = 0$; si noti inoltre che, a differenza del caso precedente, il denominatore risulta facilmente calcolabile solamente in casi notevoli e deve esser ricalcolato per ogni scelta delle componenti della distribuzione.

Tale famiglia di distribuzioni risulta molto flessibile in quanto il parametro α_0 può assumere qualsiasi valore nell'asse reale, inoltre anche la rappresentazione stocastica associata risente dell'aggiunta di tale parametro. Sia infatti Z_0 una distribuzione avente densità f_0 e $T \sim G_0$, si definisca quindi

$$S = \begin{cases} 1 & \text{se } w(Z_0) - T > -\alpha_0, \\ 0 & \text{altrimenti,} \end{cases}$$

allora la variabile $Z = (Z_0 \mid S = 1)$ ha densità definita in (2.16). Nonostante il calcolo della costante di normalizzazione possa risultare particolarmente complesso in alcuni casi, tale costruzione si riscontra in molte situazioni riguardanti dati reali. Nella pratica solitamente si sceglie $w(x) = \alpha^\top x$ mentre per convenienza matematica si sceglie sia per la variabile T che per la variabile Z_0 una distribuzione normale. Nonostante l'ampia generalità di questa famiglia, alcune problematiche sono date dalla stima dei parametri via massima verosimiglianza; per valori di $\alpha \to 0$ si ha infatti che il determinante della matrice di informazione attesa tende ad annullarsi. Ne deriva che la matrice di informazione tende ad essere quasi singolare per certi valori dei parametri portando a dei problemi nelle stime dei parametri. Nel prosieguo si vedranno casi in cui $G_0(\cdot)$ sarà diverso dalla funzione di ripartizione di una normale standardizzata.

2.6 La distribuzione normale asimmetrica estesa univariata

Il calcolo del denominatore in equazione (2.16) risulta relativamente semplice nel caso in cui le variabili Z_0 e T siano gaussiane e $w(Z_0) = \alpha Z_0$, tale configurazione prende il nome di *normale asimmetrica estesa* e viene per la prima volta definita in Azzalini (1985) e successivamente approfondita in Arnold et al. (1993), Arnold & Beaver (2000) e Azzalini & Capitanio (1999). Facendo riferimento alla densità

di probabilità definita in equazione (2.16) e definendo $\tau = \frac{\alpha_0}{\sqrt{1+\alpha^2}}$ si ottiene che la densità della variabile $Z \sim ESN(0, 1, \alpha, \tau)$ risulta pari a

$$f(z; \alpha, \tau) = \varphi(z)\Phi(\tau\sqrt{1+\alpha^2} + \alpha z)\frac{1}{\Phi(\tau)}, \quad z \in (-\infty, \infty)$$

con α e τ appartenenti ad $\mathbb{R}$. Anche in questo caso risulta possibile effettuare una trasformazione di posizione e scala, andando a definire la variabile $Y = \xi + \omega Z$ avente densità

$$f(y; \xi, \omega, \alpha, \tau) = \varphi(y - \xi; \omega)\Phi\left(\tau\sqrt{1+\alpha^2} + \alpha\frac{y-\xi}{\omega}\right)\frac{1}{\Phi(\tau)}, \quad y \in \mathbb{R}$$

indicata con $Y \sim SN(\xi, \omega^2, \alpha, \tau)$. Se il parametro α risulta pari a 0, si ha che $Y \sim N(\xi, \omega^2)$ per qualsiasi τ.

La funzione generatrice dei momenti risulta facilmente calcolabile a partire dal Lemma 3 e per il caso standard risulta essere pari a

$$M_z(t) = \exp\left(\frac{t^2}{2}\right)\Phi(\delta t + \tau)\frac{1}{\Phi(\tau)},$$

ne deriva che i primi due momenti risultano

$$\begin{aligned}\mathbb{E}[Z] &= \delta\frac{\varphi(\tau)}{\Phi(\tau)},\\ \mathbb{V}ar[Z] &= 1 - \delta^2\frac{\varphi(\tau)}{\Phi(\tau)}\left(\tau + \frac{\varphi(\tau)}{\Phi(\tau)}\right).\end{aligned}$$

La quantità $\frac{\varphi(\tau)}{\Phi(\tau)}$ risulta sempre positiva e decrescente (Figura 2.2) e viene definita *Inverse Mills Ratio*. Si noti inoltre che la funzione $\frac{\varphi(\tau)}{\Phi(\tau)}\left(\tau + \frac{\varphi(\tau)}{\Phi(\tau)}\right)$ risulta compresa nell'intervallo tra 0 e 1, ne deriva che $\mathbb{V}ar[Z] \leq 1$ per qualsiasi δ.

2.7 Rappresentazione stocastica

Similmente alla distribuzione normale asimmetrica, la distribuzione normale asimmetrica estesa può essere rappresentata come trocantura di una variabile casuale normale bivariata. Sia $(X_0, X_1) \sim N_2(0, \bar{\Omega}^*)$ dove $\bar{\Omega}^*$ ha elementi diagonali

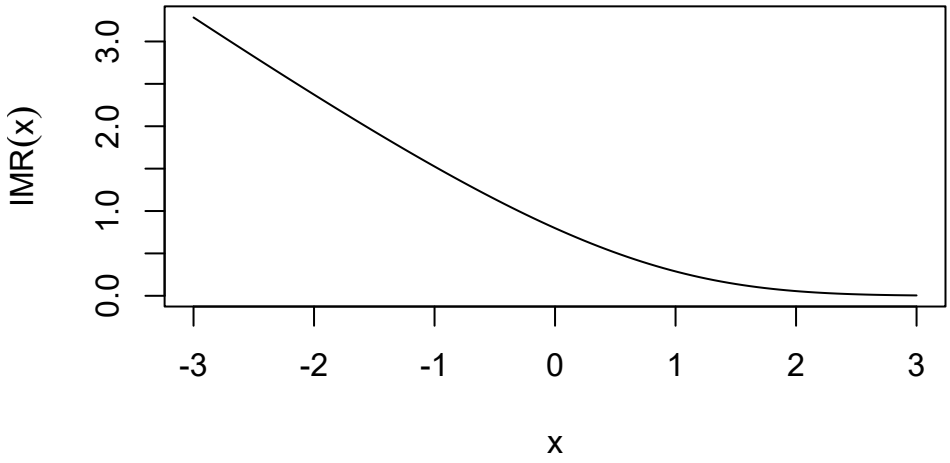

Figura 2.2: *Inverse Mills Ratio*

pari ad uno e

$$\text{corr}(X_0, X_1) = \delta = \frac{\alpha}{\sqrt{1+\alpha^2}}, \tag{2.17}$$

allora la variabile $Z = (X_0 \mid X_1 > -\tau)$ presenta una distribuzione $ESN(0, 1, \alpha, \tau)$.

Mentre il parametro di forma α porta a cambiamenti speculari sulla forma della densità a seconda che esso sia positivo o negativo, il parametro τ può comportare variazioni molto diverse a seconda che esso vari nel semiasse positivo o nel semiasse negativo. Dato il meccanismo generatore appena definito, se il troncamento della variabile tende a $-\infty$ (quindi per $\tau \to \infty$), esso diventa trascurabile. Come si nota in Figura 2.3 infatti, al crescere di τ la densità perde la sua caratteristica di asimmetria anche per valori elevati del parametro α. Al contrario se $\tau \to -\infty$ allora il condizionamento è molto forte in quanto stiamo osservando i valori sulla coda, e la variabile continua a mantenere una pronunciata asimmetria con moda verso valori più estremi (Figura 2.3 in basso).

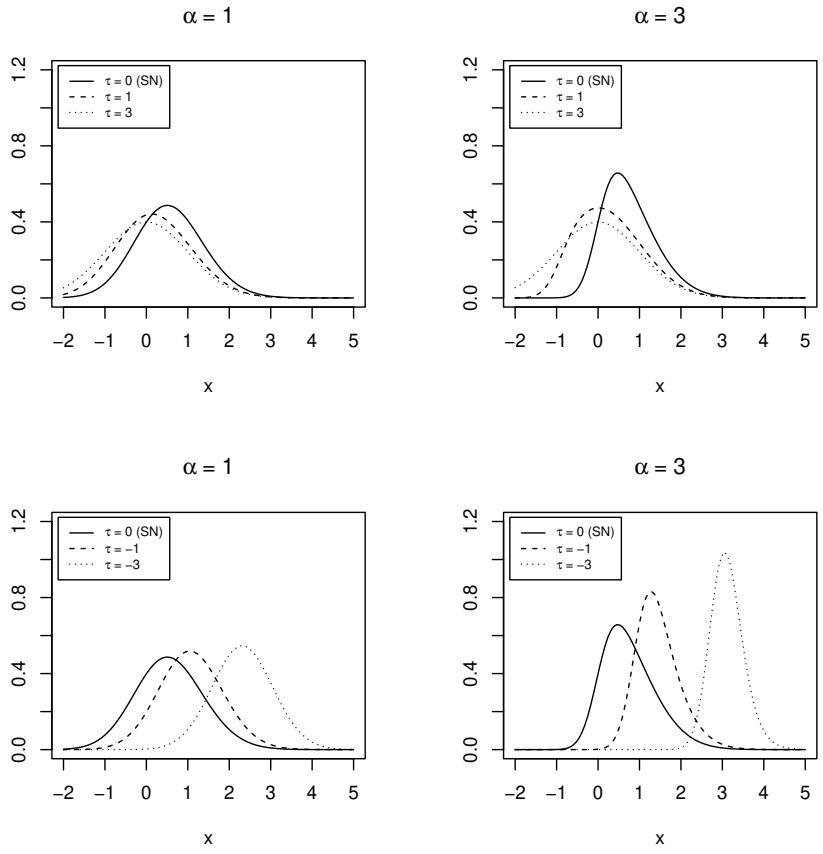

Figura 2.3: Grafico di una distribuzione $ESN(0, 1, \alpha, \tau)$ per vari valori di α e τ.

2.8 La distribuzione normale asimmetrica estesa multivariata

Risulta ora possibile andare a definire la versione multidimensionale della variabile aleatoria *extended skew normal*. Una variabile casuale d-dimensionale $Z \sim ESN_d(0, \bar{\Omega}, \alpha, \tau)$ ha densità

$$f(z; \bar{\Omega}, \alpha, \tau) = \varphi_d(z; \bar{\Omega})\Phi(\alpha_0 + \alpha^\top z)\frac{1}{\Phi(\tau)}, \quad z \in \mathbb{R}^d, \tag{2.18}$$

dove α è un vettore d-dimensionale, $\bar{\Omega}$ è una matrice di correlazione di dimensioni $d \times d$ e $\alpha_0 = \tau(1 + \alpha^\top \bar{\Omega}\alpha)^{1/2}$. Similmente al caso univariato l'effetto di τ svanisce se il vettore α è nullo.

Sia Z una variabile aleatoria con densità in (2.18) allora la variabile $Y = \xi + \omega Z$ ha densità che risulta

$$f(y; \xi, \Omega, \alpha, \tau) = \varphi_d(y - \xi; \Omega)\Phi(\alpha_0 + \alpha^\top \omega^{-1}(y - \xi))\frac{1}{\Phi(\tau)}, \quad y \in \mathbb{R}^d \tag{2.19}$$

con la stessa notazione di (2.5).

Utilizzando il Lemma 4 si ottiene che la funzione generatrice dei momenti per Y risulta pari a

$$M_y(t) = \exp\left(t^\top \xi + \frac{1}{2}t^\top \Omega t\right)\Phi(\tau + \delta^\top \omega t)\Phi(\tau)^{-1}, \tag{2.20}$$

con

$$\delta = (1 + \alpha^\top \bar{\Omega}\alpha)^{-1/2}\bar{\Omega}\alpha$$

vettore di correlazione. Anche in questo caso il parametro τ fa assumere alla distribuzione forme diverse a seconda che esso si trovi nell'asse positivo o nell'asse negativo (Figura 2.4).

2.8.1 Rappresentazione stocastica multivariata

Risulta naturale estendere la rappresentazione stocastica della SN alla variabile SN estesa. Sia quindi (X_0, X_1) il vettore aleatorio definito in (2.6) allora, per ogni

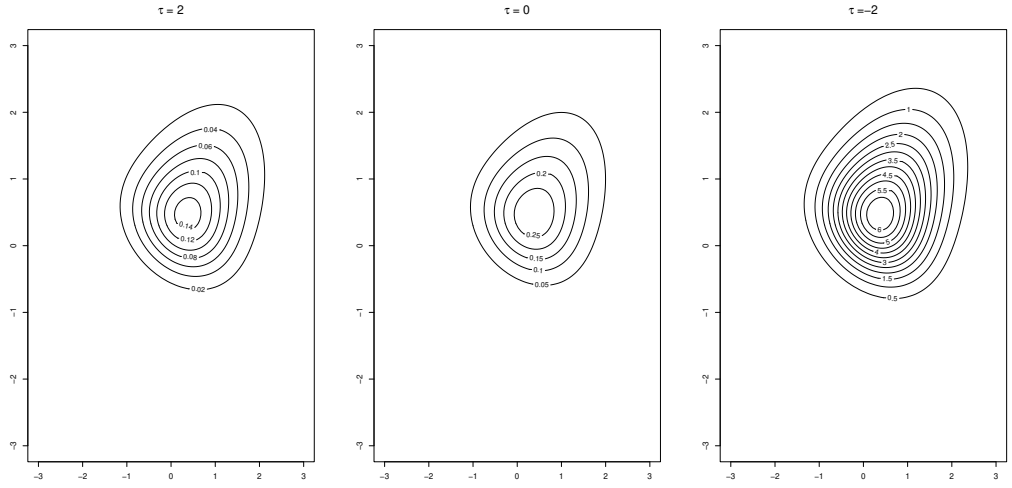

Figura 2.4: Grafico di densità di una distribuzione $ESN(0, \bar{\Omega}, \alpha, \tau)$ per vari valori di τ, con $\alpha = (1, 2)$ e $\bar{\omega}_{xw} = 0.5$.

$\tau \in \mathbb{R}$, si ha

$$Z = (X_0 \mid X_1 > -\tau) \sim ESN_d(0, \bar{\Omega}, \alpha, \tau) \tag{2.21}$$

con α definito in (2.7).

Il meccanismo generatore per ottenere la variabile $Y \sim ESN(\xi, \Omega, \alpha, \tau)$ risulta praticamente coincidente; sia

$$\begin{pmatrix} X_0^* \\ X_1^* \end{pmatrix} \sim N_{d+1}(\xi^*, \Omega^*), \quad \text{con } \Omega^* = \omega^* \bar{\Omega}^* \omega^* = \omega^* \begin{pmatrix} \bar{\Omega} & \delta \\ \delta^\top & 1 \end{pmatrix} \omega^*,$$

con $\bar{\Omega}^*$ matrice di correlazione, $\xi^* = (\xi, \xi_1)^\top$ e $\omega^* = (I_{d+1} \odot \Omega^*)^{1/2} = \text{diag}(\omega, \omega_1)$ matrice diagonale di dimensioni $(d+1) \times (d+1)$. Ne deriva che marginalmente la variabile X_1^* risulta essere normale di media ξ_1 e varianza ω_1^2 mentre $X_0^* \sim N_d(\xi, \Omega)$. La variabile Y si ottiene condizionando la variabile X_0^* ai valori di X_1^* superiori alla soglia $\xi_1 - \omega_1 \tau$, utilizzando le trasformazioni di posizione e scala infatti si ottiene

$$\begin{aligned} (X_0^* \mid X_1^* > \xi_1 - \omega_1 \tau) &= (X_0^* \mid \xi_1 + \omega_1 X_1 > \xi_1 - \omega_1 \tau) \\ &= (X_0^* \mid X_1 > -\tau) \\ &= (\xi + \omega X_0 \mid X_1 > -\tau) \end{aligned}$$

che, utilizzando il risultato ottenuto in equazione (2.21), risulta distribuirsi come una *extended skew normal* con densità definita in (2.19).

2.8.2 Distribuzione marginale e condizionata

Una delle proprietà più interessanti della distribuzione normale asimmetrica estesa multivariata è la chiusura rispetto sia al condizionamento che alla marginalizzazione come mostrato in Azzalini & Dalla Valle (1996) e successivamente in Azzalini & Capitanio (1999).

Partizionando la variabile $Y^\top = (X^\top, W^\top)$ e calcolando la (2.20) in $t^\top = (s^\top, 0)$, con $dim(X) = dim(s) = h$, si ottiene

$$X \sim ESN(\xi_x, \Omega_{xx}, \alpha_{x(w)}, \tau) \tag{2.22}$$

dove i primi tre elementi sono definiti rispettivamente in (2.12) e in (2.15), mentre il parametro τ rimane invariato rispetto alla distribuzione congiunta.

Un'altra importante proprietà della distribuzione è la chiusura rispetto al condizionamento. Sia infatti $Y^\top = (X^\top, W^\top)$, dove X ha dimensione h e sia di interesse studiare la distribuzione di W dato $X = x$. Nel caso in cui $Y \sim N_d(\xi, \Omega)$ allora i parametri di W condizionatamente ai valori assunti dalla variabile X risultano

$$\begin{aligned} \xi_{w.x} &= \xi_w + \Omega_{wx}\Omega_{xx}^{-1}(x - \xi_x), \\ \Omega_{ww.x} &= \Omega_{ww} - \Omega_{wx}\Omega_{xx}^{-1}\Omega_{xw}, \end{aligned}$$

essendo la distribuzione normale asimmetrica estesa una estensione della distribuzione normale, tali parametri compaiono anche nella distribuzione di $(W \mid X = x)$. La legge condizionata di W dato $X = x$ risulta

$$\varphi_{d-h}(w - \xi_{w.x}; \Omega_{ww.x}) \frac{\Phi\big(\alpha_0' + \alpha_w^\top \omega_w^{-1}(w - \xi_{w.x})\big)}{\Phi(\tau_{w.x})}, \quad y \in \mathbb{R}^{d-h}$$

dove

$$\begin{aligned}
\tau_{w.x} &= \tau(1+\alpha_{x(w)}^{\top}\bar{\Omega}_{xx}\alpha_{x(w)})^{1/2} + \alpha_{x(w)}^{\top}\omega_x^{-1}(x-\xi_x),\\
\alpha_0^{'} &= \tau_{w.x}(1+\alpha_{w.x}^{\top}\Omega_{ww.x}^{-1}\alpha_{w.x})^{1/2},\\
\alpha_{w.x} &= \omega_{ww.x}\omega_w^{-1}\alpha_w,\\
\omega_{ww.x} &= (\Omega_{ww.x}\odot I_{d-h})^{1/2}
\end{aligned} \tag{2.23}$$

riaggiustando i termini, ne deriva che

$$(W \mid X=x) \sim ESN_{d-h}(\xi_{w.x}, \Omega_{ww.x}, \alpha_{w.x}, \tau_{w.x})$$

che conferma quanto detto in precedenza.

Si noti che se $\alpha_w = 0$ allora anche $\alpha_{w.x}$ avrà valore pari a 0, ciò comporta che la distribuzione di W condizionata ad $X = x$ risulta gaussiana. Tale fatto può esser generalizzato, sia infatti Y una distribuzione normale asimmetrica allora se l'r-esimo elemento di α è nullo, si ha che l'r-esima componente di Y ha distribuzione gaussiana se condizionata al valore delle altre componenti. Partendo da queste considerazioni, risulta possibile studiare l'indipendenza condizionata a coppie (Capitanio et al., 2003).

Lemma 5 *Sia* $Y \sim ESN_d(\xi, \Omega, \alpha, \tau)$ *allora*

$$Y_i \perp\!\!\!\perp Y_j \mid Y_{-\{ij\}}$$

se e solo se le seguenti condizioni vengono rispettate:

1. Ω^{ij}*=0,*
2. $\alpha_i\alpha_j = 0$

dove Ω^{ij} *indica l'elemento in posizione* (i,j) *della matrice* Ω^{-1}.

Tale risultato risulta fondamentale per lo sviluppo di modelli grafici utilizzando la distribuzione ESN.

Capitolo 3

Relazioni tra parametri condizionati e marginali

Nel Capitolo 1 viene posta l'attenzione sulla *mediation analysis* e sulle metodologie utilizzate in questo contesto. Come già sottolineato, mentre nel caso dei modelli lineari per dati con risposta e mediatore continui le relazioni risultano relativamente semplici, nella situazione di dati con risposta dicotomica le relazioni si fanno complesse e sorgono alcune problematiche derivanti dalla natura della variabile di interesse. In questo Capitolo verrà posta l'attenzione sulla ricerca di formulazioni relativamente semplici ed interpretabili che leghino i parametri marginali e condizionati nei modelli di regressione logistica.

3.1 Regressione logistica con mediatore binario e trattamento continuo

Esistono situazioni dove sia la risposta che il mediatore hanno scala binaria, si pensi ad esempio al caso in cui sia di interesse studiare l'uso di droghe da parte degli adolescenti e la variabile mediatrice sia rappresentata dal consumo di alcolici o di sigarette mentre la covariata d'interesse risulta essere il reddito. Il lavoro di Stanghellini & Doretti (2019) cerca di scomporre l'effetto totale di X su Y in scala log-rapporto di quote nella situazione in cui Y è una risposta binaria, W un mediatore binario mentre X è un trattamento continuo. In particolare si postula

un modello di regressione logistica per Y dati i valori di X e W,

$$\log \frac{\mathbb{P}(Y = 1 \mid X = x, W = w)}{\mathbb{P}(Y = 0 \mid X = x, W = w)} = \beta_0 + \beta_x x + \beta_w w + \beta_{xw} xw, \tag{3.1}$$

e un modello di regressione logistica per W dato X,

$$\log \frac{\mathbb{P}(W = 1 \mid X = x)}{\mathbb{P}(W = 0 \mid X = x)} = \theta_0 + \theta_x x. \tag{3.2}$$

Si noti che il parametro β_{xw} permette l'interazione tra le covariate X e W. L'effetto marginale di X su Y in scala logit è definito dalla derivata

$$\beta(x) = \frac{d}{dx} \log \frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = 0 \mid X = x)}, \tag{3.3}$$

dove la notazione $\beta(x)$ è dovuta al fatto che l'effetto marginale varia con x in maniera non lineare, infatti se vale il modello in (3.1) allora il logit marginale presente nella parte destra dell'equazione (3.3) non è lineare in x (Lin et al., 1998).

In particolare, viene dimostrato che la quantità $\beta(x)$ risulta essere pari a

$$\begin{aligned} \beta(x) =& \beta_x\{1 - \Delta_y(x)\Delta_w(x)\} + \theta_x \Delta_w(x) \\ &+ \beta_{xw}\{\mathbb{P}(W = 1 \mid Y = 1, X = x) - \Delta_w(x)\mathbb{P}(Y = 1 \mid W = 1, X = x)\}, \end{aligned} \tag{3.4}$$

dove

$$\Delta_y(x) = \mathbb{P}(Y = 1 \mid W = 1, X = x) - \mathbb{P}(Y = 1 \mid W = 0, X = x),$$

e

$$\Delta_w(x) = \mathbb{P}(W = 1 \mid Y = 1, X = x) - \mathbb{P}(W = 1 \mid Y = 0, X = x).$$

La formula definita in (3.4) presenta il vantaggio di rendere esplicito il modo in cui i parametri delle distribuzioni condizionate si combinano per formare l'effetto marginale di X su Y in scala log-rapporto di quote. Tale formulazione può essere vista come una generalizzazione della *scomposizione di Cochran* applicata al caso della regressione logistica con mediatore binario, in quanto la relazione dipende dalla somma di componenti che svaniscono se alcuni dei parametri dei modelli in

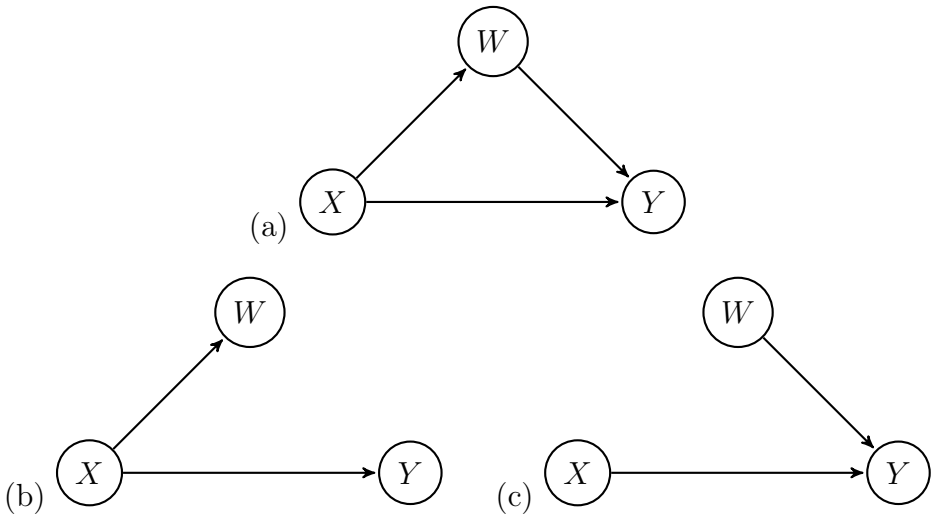

Figura 3.1: Grafico relazione quando (a) non possono essere effettuate assunzioni di indipendenza, (b) $W \perp\!\!\!\perp Y \mid X$ e (c) $X \perp\!\!\!\perp W$

(3.1) e (3.2) sono nulli.

Risulta a questo punto semplice definire i vari sottocasi derivanti dall'utilizzo della formula (3.4), ad esempio, se non è presente l'interazione tra i regressori deriva che

$$\beta(x) = \beta_x\{1 - \Delta_y(x)\Delta_w(x)\} + \theta_x\Delta_w(x).$$

Se anche il coefficiente $\beta_w = 0$, allora W e Y risultano condizionatamente indipendenti dato X, in simboli $W \perp\!\!\!\perp Y \mid X$, ne consegue che anche $\Delta_y(x)$ e $\Delta_w(x)$ sono nulli e si arriva al risultato

$$\beta(x) = \beta_x,$$

in linea con quanto dimostrato in Xie et al. (2008). Nel caso in cui $\theta_x = \beta_{xw} = 0$, quindi $W \perp\!\!\!\perp X$, c'è comunque un effetto sul parametro dovuto al noto risultato di non collassabilità del parametro del modello di regressione logistica, dimostrato ad esempio in (Neuhaus & Jewell, 1993), e la relazione risulta

$$\beta(x) = \beta_x\{1 - \Delta_y(x)\Delta_w(x)\},$$

ne deriva che $|\beta(x)| \leq |\beta_x|$, in quanto la quantità tra parentesi risulta essere compresa nell'intervallo chiuso tra 0 e 1. I tre processi generatori dei dati appena descritti possono essere rappresentati attraverso il grafo aciclico rappresentato in

Figura 3.1, dove la mancanza di una delle frecce corrisponde ad una delle assunzioni di indipendenza sopra delineate.

3.2 Regressione logistica con mediatore e trattamento continui

Nel prosieguo si farà riferimento al caso in cui sia X che W sono variabili continue mentre la risposta Y ha natura binaria. In particolare si farà riferimento ad uno specifico meccanismo generatore dei dati comprendente tre variabili aleatorie continue (X, W, Y^*) congiuntamente gaussiane, dove la variabile risposta rappresenta una dicotomizzazione della variabile latente Y^*. Anche in questo caso si cercherà una relazione che leghi i parametri marginali e condizionati e che permetta una divisione dell'effetto totale in effetti diretti ed indiretti che sia di facile interpretazione. Gli sviluppi seguenti, dalle nostre conoscenze, non sembrerebbero essere presenti in letteratura.

3.3 Descrizione del *Data Generating Process*

Sia Y la variabile binaria ottenuta dalla dicotomizzazione della variabile latente $Y^* \sim N(\mu_y, \omega_{yy})$,

$$Y = \begin{cases} 1 & \text{se } \frac{Y^* - \mu_y}{\sqrt{\omega_{yy}}} > -\tau, \\ 0 & \text{altrimenti,} \end{cases} \tag{3.5}$$

dove τ è un valore nell'asse reale.

Si definisca inoltre la distribuzione del vettore aleatorio $(X, W, Y^*)^\top$, dove X e W rappresentano le covariate di interesse utili nello spiegare la risposta Y, in particolare

$$\begin{pmatrix} X \\ W \\ Y^* \end{pmatrix} \sim N_3(\xi^*, \Omega^*), \quad \Omega^* = \omega^* \bar{\Omega}^* \omega^* = \omega^* \begin{pmatrix} \bar{\Omega} & \delta \\ \delta^\top & 1 \end{pmatrix} \omega^*, \tag{3.6}$$

con $\xi^* = (\mu_x, \mu_w, \mu_y)^\top$, $\bar{\Omega}^*$ matrice di correlazione 3×3, $\omega^* = \text{diag}(\sqrt{\omega_{xx}}, \sqrt{\omega_{ww}}, \sqrt{\omega_{yy}})$ matrice diagonale contenente le deviazioni standard degli elementi del vettore alea-

torio mentre $\delta^\top = (\delta_x, \delta_w)^\top$ contiene le correlazioni tra la variabile risposta e le variabili indipendenti. Tale distribuzione congiunta permette di ottenere facilmente la distribuzione condizionata delle covariate rispetto al valore della risposta. Come dimostrato nel Paragrafo 2.8.1, sfruttando la rappresentazione stocastica (2.21), si ha

$$\begin{aligned} (X, W|Y = 1) &\sim ESN_2(\xi, \Omega, \alpha, \tau), \\ (X, W|Y = 0) &\sim ESN_2(\xi, \Omega, -\alpha, -\tau), \end{aligned} \tag{3.7}$$

dove i parametri sono definiti come segue,

$$\begin{aligned} \xi &= (\mu_x, \mu_w)^\top, \\ \omega &= \text{diag}(\sqrt{\omega_{xx}}, \sqrt{\omega_{ww}}), \\ \Omega &= \omega\bar{\Omega}\omega, \\ \bar{\Omega} &= \begin{pmatrix} 1 & \bar{\omega}_{xw} \\ \bar{\omega}_{xw} & 1 \end{pmatrix}, \\ \alpha &= (\alpha_x, \alpha_w)^\top = (1 - \delta^\top \bar{\Omega}^{-1}\delta)^{-1/2}\bar{\Omega}^{-1}\delta. \end{aligned} \tag{3.8}$$

Tale processo generatore dei dati risulta molto flessibile ed utile nel descrivere fenomeni nei quali la risposta viene definita tramite discretizzazione di una sottostante variabile latente. Il parametro $\bar{\omega}_{xw}$ misura la dipendenza tra le due covariate, in particolare, X e W risultano correlate nel caso in cui $\bar{\omega}_{xw}$ sia diverso da 0. Nel caso in cui il vettore δ sia pari a 0 allora entrambe le distribuzioni condizionate risultano normali di media ξ e matrice di varianze e covarianze pari ad Ω, in quanto il vettore α risulta nullo.

Come si nota dalla Figura 3.2, all'aumentare del valore assoluto dei parametri contenuti nel vettore α, le classi risultano essere più separate; mentre il parametro τ agisce in termini di forma e di variabilità sulle distribuzioni delle variabili *normali asimmetriche estese*. Si noti che se $\tau = 0$ allora le due classi risultano perfettamente bilanciate e la forma della densità risulta essere speculare per le due classi.

3.3.1 Derivazione del logit condizionato

Risulta ora di interesse studiare la forma funzionale del logit condizionatamente ai valori di ambedue i regressori. Sfruttando risultati base di probabilità si ottiene

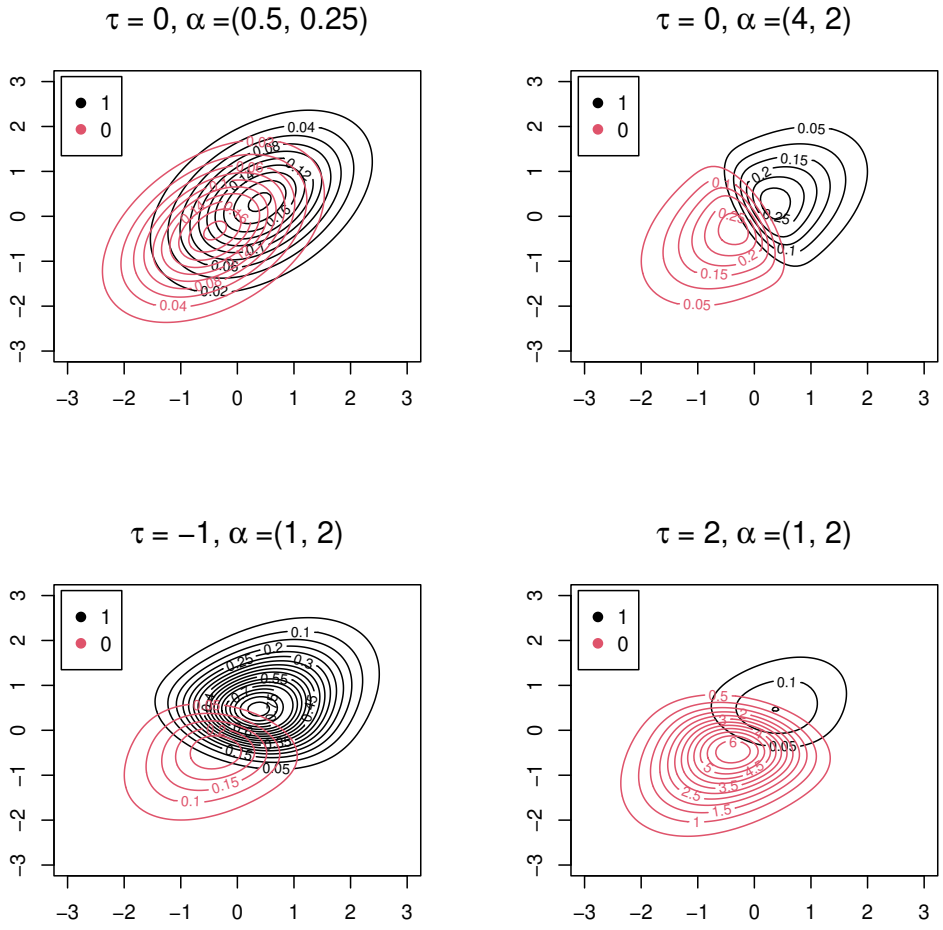

Figura 3.2: Distribuzione di $(X, W \mid Y = y)$ al variare di τ e α, con $\bar{\omega}_{xw} = 0.5$ ed $\omega = I$. In nero la distribuzione rispetto ad $y = 1$, in rosso la distribuzione rispetto ad $y = 0$.

$$\begin{aligned}\log \frac{\mathbb{P}(Y=1|X=x,W=w)}{\mathbb{P}(Y=0|X=x,W=w)} &= \log \frac{f(x,w \mid Y=1)\mathbb{P}(Y=1)}{f(x,w)} + \\ &\quad - \log \frac{f(x,w \mid Y=0)\mathbb{P}(Y=0)}{f(x,w)} \\ &= \log \frac{f(x,w|Y=1)}{f(x,w|Y=0)} + \log \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}.\end{aligned} \tag{3.9}$$

Si noti come l'ultimo addendo non dipende né da x né da w e dato il processo generatore dei dati descritto nella Sezione precedente si ottiene che la probabilità marginale per la prima classe risulta $\mathbb{P}(Y = 1) = \Phi(\tau)$. Sfruttando i risultati ottenuti in precedenza si ha

$$\begin{aligned}\log \frac{\mathbb{P}(Y=1|X=x,W=w)}{\mathbb{P}(Y=0|X=x,W=w)} &= \log \frac{\varphi_2(z-\xi;\Omega)\Phi(\alpha_0+\alpha^\top\omega^{-1}(z-\xi))\Phi(\tau)^{-1}}{\varphi_2(z-\xi;\Omega)\Phi(-\alpha_0-\alpha^\top\omega^{-1}(z-\xi))\Phi(-\tau)^{-1}} + \\ &\quad + \log \frac{\Phi(\tau)}{1-\Phi(\tau)} \\ &= \log \frac{\Phi(\alpha_0+\alpha^\top\omega^{-1}(z-\xi))}{1-\Phi(\alpha_0+\alpha^\top\omega^{-1}(z-\xi))} \\ &= \log \frac{\Phi\left(\alpha_0-\alpha^\top\omega^{-1}\xi+\frac{\alpha_x}{\sqrt{\omega_{xx}}}x+\frac{\alpha_w}{\sqrt{\omega_{ww}}}w\right)}{1-\Phi\left(\alpha_0-\alpha^\top\omega^{-1}\xi+\frac{\alpha_x}{\sqrt{\omega_{xx}}}x+\frac{\alpha_w}{\sqrt{\omega_{ww}}}w\right)}\end{aligned} \tag{3.10}$$

dove $z = (x, w)$ e $\alpha_0 = \tau(1 + \alpha^\top\bar{\Omega}\alpha)^{1/2}$. Si noti come tale forma funzionale sia il log-rapporto di quote tra due funzioni contenenti le variabili d'interesse al loro interno, e risulti quindi non lineare. Tale quantità risulta non dipendere da x e w solamente se il vettore α risulta nullo, cioè se le variabili X e W sono indipendenti dalla variabile latente Y^*.

3.4 Derivazione del logit marginale

Si calcola ora il logit marginale rispetto alla variabile X per ottenere la quantità definita in (3.3). A partire da risultati base del calcolo delle probabilità verranno

in seguito presentati due modi per derivare il logit marginale.

3.4.1 Primo metodo

Il primo metodo si ottiene sviluppando l'equazione definita in (3.9), ottenendo la seguente identità

$$\log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \log \frac{f(x \mid Y=1)}{f(x \mid Y=0)} + \log \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}, \tag{3.11}$$

dove l'ultimo membro a destra del segno di uguaglianza non risulta dipendere da x.

Sfruttando la proprietà di chiusura rispetto alla marginalizzazione della distribuzione *Extended Skew Normal* definita nel Paragrafo 2.8.2, si ottiene che la distribuzione marginale di X dato il valore di $Y = y$ rimane ESN con i seguenti parametri

$$\begin{aligned} (X \mid Y=1) &\sim ESN(\xi_x, \omega_{xx}, \alpha_{x(w)}, \tau), \\ (X \mid Y=0) &\sim ESN(\xi_x, \omega_{xx}, -\alpha_{x(w)}, -\tau), \end{aligned} \tag{3.12}$$

dove $\alpha_{x(w)} = (\alpha_x + \bar{\omega}_{xw}\alpha_w)(1+\alpha_w^2(1-\bar{\omega}_{xw}^2))^{-1/2}$, mentre i parametri α_x, α_w e $\bar{\omega}_{xw}$ sono definiti in (3.8). Risulta ora possibile calcolare la forma funzionale del logit marginale, sviluppando l'equazione (3.11) si ottiene

$$\begin{aligned} \log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} &= \log \frac{\varphi\big(x-\xi_x;\omega_{xx}\big)\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2}+\alpha_{x(w)}\frac{x-\xi_x}{\sqrt{\omega_{xx}}}\big)\Phi(\tau)^{-1}}{\varphi\big(x-\xi_x;\omega_{xx}\big)\Phi\big(-\tau\sqrt{1+\alpha_{x(w)}^2}-\alpha_{x(w)}\frac{x-\xi_x}{\sqrt{\omega_{xx}}}\big)\Phi(-\tau)^{-1}} + \\ &\quad + \log \frac{\Phi(\tau)}{\Phi(-\tau)} \\ &= \log \frac{\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2}+\alpha_{x(w)}\frac{x-\xi_x}{\sqrt{\omega_{xx}}}\big)}{1-\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2}+\alpha_{x(w)}\frac{x-\xi_x}{\sqrt{\omega_{xx}}}\big)} \\ &= \log \frac{\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2}-\frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}\xi_x+\frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}x\big)}{1-\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2}-\frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}\xi_x+\frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}x\big)} \end{aligned} \tag{3.13}$$

La forma funzionale risulta simile all'equazione (3.10) in quanto le distribuzioni condizionate al valore della risposta risultano *normali asimmetriche estese* in ambedue i casi. Tale quantità non risulta dipendere da x solamente se il coefficiente $\alpha_{x(w)}$ risulta nullo. Assumendo la nullità di $\alpha_{x(w)}$ si ottiene infatti

$$\log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \log \frac{\Phi(\tau)}{1-\Phi(\tau)},$$

e sfruttando le proprietà definite nel Paragrafo 2.6, si ha che se $\alpha_{x(w)} = 0$ allora le distribuzioni di X dato Y risultano gaussiane per qualsiasi valore di τ.

3.4.2 Secondo metodo

Alternativamente si può partire dalla seguente relazione derivante da risultati base del calcolo delle probabilità

$$\log \frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} = \log \frac{f(w|Y=0,X=x)}{f(w|Y=1,X=x)} + \log \frac{\mathbb{P}(Y=1|X=x,W=w)}{\mathbb{P}(Y=0|X=x,W=w)},$$

dove il secondo membro a destra del segno di uguale coincide con l'equazione (3.10).

Risulta ora necessario ricavare la distribuzione W condizionatamente ai valori assunti dalle variabili X ed Y. Sfruttando le identità definite nel Paragrafo 2.8.2 si ottiene che

$$\begin{aligned}(W|X=x,Y=1) &\sim ESN(\xi_{w.x}, \omega_{ww.x}, \alpha_{w.x}, \tau_{w.x}),\\ (W|X=x,Y=0) &\sim ESN(\xi_{w.x}, \omega_{ww.x}, -\alpha_{w.x}, -\tau_{w.x}),\end{aligned} \tag{3.14}$$

dove i parametri sono definiti come

$$\begin{aligned}\xi_{w.x} &= \xi_w + \omega_{wx}\omega_{xx}^{-1}(x-\xi_x),\\ \omega_{ww.x} &= \omega_{ww} - \omega_{xw}^2\omega_{xx}^{-1},\\ \alpha_{w.x} &= (\omega_{ww.x}^{1/2})\omega_{ww}^{-1/2}\alpha_w,\\ \tau_{w.x} &= \tau\sqrt{1+\alpha_{x(w)}^2} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x).\end{aligned} \tag{3.15}$$

Risulta ora possibile sviluppare l'identità sopra definita

$$\log \frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} = \log \frac{\Phi\big(-\tau_{w.x}\sqrt{1+\alpha_{w.x}^2} - \alpha_{w.x}\frac{w-\xi_{w.x}}{\sqrt{\omega_{ww.x}}}\big)\Phi(-\tau_{w.x})^{-1}}{\Phi\big(\tau_{w.x}\sqrt{1+\alpha_{w.x}^2} + \alpha_{w.x}\frac{w-\xi_{w.x}}{\sqrt{\omega_{ww.x}}}\big)\Phi(\tau_{w.x})^{-1}} + \\ + \log \frac{\varphi(w-\xi_{w.x};\omega_{ww.x})}{\varphi(w-\xi_{w.x};\omega_{ww.x})} + \log \frac{\Phi(\alpha_0+\alpha^\top\omega^{-1}(z-\xi))}{\Phi(-\alpha_0-\alpha^\top\omega^{-1}(z-\xi))}$$

Si noti che la quantità $\tau_{w.x}\sqrt{1+\alpha_{w.x}^2} + \alpha_{w.x}\frac{w-\xi_{w.x}}{\sqrt{\omega_{ww.x}}}$ presente nel primo addendo all'interno di $\Phi(\cdot)$, può essere riscritta come

$$\begin{aligned}
\tau_{w.x}\sqrt{1+\alpha_{w.x}^2} + \alpha_{w.x}\frac{w-\xi_{w.x}}{\sqrt{\omega_{ww.x}}} =& \tau\sqrt{1+\alpha_{x(w)}^2} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x) + \\
&+ \alpha_w\frac{w-\xi_{w.x}}{\sqrt{\omega_{ww}}} \\
=& \tau\sqrt{(1+\alpha_{x(w)}^2)(1+\alpha_{w.x}^2)} + \\
&+ \left[\frac{\alpha_{x(w)}\sqrt{1+\alpha_{w.x}^2} - \alpha_w\bar{\omega}_{xw}}{\sqrt{\omega_{xx}}}\right](x-\xi_x) + \\
&+ \alpha_w\frac{w-\xi_w}{\sqrt{\omega_{ww}}} \\
=& \tau\sqrt{1+\alpha^\top\bar{\Omega}\alpha} + \alpha_x\frac{x-\xi_x}{\sqrt{\omega_{xx}}} + \alpha_w\frac{w-\xi_w}{\sqrt{\omega_{ww}}} \\
=& \alpha_0 + \alpha^\top\omega^{-1}(z-\xi).
\end{aligned}$$

Risulta a questo punto possibile applicare la seguente semplificazione

$$\log \frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} = \log \frac{\Phi(\tau_{w.x})}{\Phi(-\tau_{w.x})},$$

ed esplicitando il parametro $\tau_{w.x}$ rispetto alla variabile x come fatto in (3.15) si ottiene

$$\log \frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} = \log \frac{\Phi\big(\tau\sqrt{1+\alpha_{x(w)}^2} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)}{\Phi\big(-\tau\sqrt{1+\alpha_{x(w)}^2} - \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)}, \tag{3.16}$$

che coincide con l'equazione definita in (3.13). Tale derivazione mette tuttavia in risalto i parametri del logit marginale con i parametri della distribuzione condizio-

nata della variabile casuale W.

3.5 Calcolo di $\beta(x)$

Risulta a questo punto possibile calcolare la quantità $\beta(x)$ definita in equazione (3.3). Si noti che anche in questo caso il logit marginale non è lineare in x, e quindi la notazione $\beta(x)$ indica che l'effetto marginale varia non linearmente in x. Derivando il risultato ottenuto in (3.13), si ottiene

$$\begin{aligned}\beta(x) =& \frac{d}{dx}\log\frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} \\ =& \frac{d}{dx}\Bigg\{ \log\Phi\left(\tau\sqrt{1+\alpha^2_{x(w)}} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\right) + \\ & - \log\Phi\left(-\tau\sqrt{1+\alpha^2_{x(w)}} - \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\right)\Bigg\}\end{aligned}$$

A questo punto risulta necessario applicare il seguente risultato per calcolare la derivata rispetto ad x.

Lemma 6 *Sia $f(x)$ una funzione in x definita tramite*

$$f(x) = \log\Phi(a+bx) - \log\Phi(-a-bx),$$

con $a, b \in \mathbb{R}$, allora

$$\frac{d}{dx}f(x) = b\left(\frac{\varphi(a+bx)}{\Phi(a+bx)} + \frac{\varphi(-a-bx)}{\Phi(-a-bx)}\right).$$

A questo punto risulta possibile calcolare l'espressione per $\beta(x)$,

$$\begin{aligned}\beta(x) =& \alpha_{x(w)}\omega_{xx}^{-1/2}\Bigg[\frac{\varphi\big(\tau\sqrt{1+\alpha^2_{x(w)}} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)}{\Phi\big(\tau\sqrt{1+\alpha^2_{x(w)}} + \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)} + \\ & + \frac{\varphi\big(-\tau\sqrt{1+\alpha^2_{x(w)}} - \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)}{\Phi\big(-\tau\sqrt{1+\alpha^2_{x(w)}} - \alpha_{x(w)}\omega_{xx}^{-1/2}(x-\xi_x)\big)}\Bigg]\end{aligned} \tag{3.17}$$

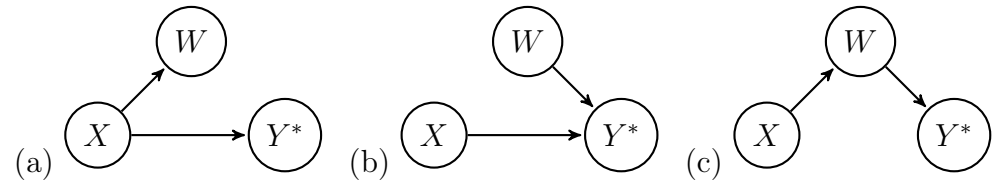

Figura 3.3: Grafico delle relazioni (a) $W \perp\!\!\!\perp Y^* \mid X$ e (b) $X \perp\!\!\!\perp W$ e (c) $X \perp\!\!\!\perp Y^* \mid W$.

Sfruttando il risultato in equazione (3.16), risulta equivalente scrivere

$$\beta(x) = \frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right] \tag{3.18}$$

dove la notazione $\tau_{w.x}(x)$ sottolinea la dipendenza della quantità dalla variabile x mentre il parametro $\alpha_{x(w)}$ è definito da,

$$\alpha_{x(w)} = \frac{\alpha_x + \bar{\omega}_{xw}\alpha_w}{\sqrt{1 + \alpha_w^2(1 - \bar{\omega}_{xw}^2)}}.$$

Tale risultato risulta abbastanza complesso in quanto coinvolge la somma tra due *Inverse Mills Ratio*. Risulta tuttavia possibile scrivere tale relazione come somma di due quantità, infatti

$$\begin{aligned}\beta(x) =& \frac{1}{\sqrt{\omega_{xx}}}\frac{\alpha_x + \bar{\omega}_{xw}\alpha_w}{\sqrt{1 + \alpha_w^2(1 - \bar{\omega}_{xw}^2)}}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right] \\ =& \frac{\alpha_x}{k}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right] + \frac{\alpha_w\bar{\omega}_{xw}}{k}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right],\end{aligned} \tag{3.19}$$

dove $k = \sqrt{\omega_{xx}}\sqrt{1 + \alpha_w^2(1 - \bar{\omega}_{xw}^2)}$, mentre la quantità tra parentesi quadre risulta essere sempre positiva in quanto somma di rapporti tra quantità positive.

3.5.1 Relazioni tra variabili

Al fine di ricavare le relazioni che legano effetti marginali e condizionati con le relazioni presenti nel vettore aleatorio (X, W, Y^*) risulta necessario espandere i parametri presenti nel vettore α. Sviluppando l'equazione (2.7) facendo riferimento

al processo generatore sviluppato in (3.5) e (3.6), si ottiene

$$\begin{aligned}\alpha &= \frac{1}{\sqrt{1-\delta^\top\bar{\Omega}^{-1}\delta}}\bar{\Omega}^{-1}\delta \\ &= \frac{1}{\sqrt{1-\delta^\top\bar{\Omega}^{-1}\delta}}\begin{pmatrix}1 & -\bar{\omega}_{xw}\\ -\bar{\omega}_{xw} & 1\end{pmatrix}\begin{pmatrix}\delta_x\\ \delta_w\end{pmatrix} \\ &= \frac{1}{\sqrt{1-\delta^\top\bar{\Omega}^{-1}\delta}}\begin{pmatrix}\delta_x-\delta_w\bar{\omega}_{xw}\\ \delta_w-\delta_x\bar{\omega}_{xw}\end{pmatrix}\end{aligned}$$

I parametri α_x ed α_w risultano quindi rispettivamente pari a

$$\begin{aligned}\alpha_x &= \frac{\delta_x-\delta_w\bar{\omega}_{xw}}{\sqrt{1-(\delta_x^2-2\delta_x\delta_w\bar{\omega}_{xw}+\delta_w^2)}},\\ \alpha_w &= \frac{\delta_w-\delta_x\bar{\omega}_{xw}}{\sqrt{1-(\delta_x^2-2\delta_x\delta_w\bar{\omega}_{xw}+\delta_w^2)}}.\end{aligned}$$

A questo punto risulta di interesse notare che i parametri α_x ed α_w sono proporzionali rispettivamente alla correlazione parziale tra X ed Y^* dato W e tra W ed Y^* dato X. Sfruttando le proprietà descritte in Appendice A.2, si ottiene infatti che

$$\begin{aligned}\rho_{X\,Y^*\cdot W} &= \frac{\delta_x-\delta_w\bar{\omega}_{xw}}{\sqrt{1-\delta_w^2}\sqrt{1-\bar{\omega}_{xw}^2}},\\ \rho_{W\,Y^*\cdot X} &= \frac{\delta_w-\delta_x\bar{\omega}_{xw}}{\sqrt{1-\delta_x^2}\sqrt{1-\bar{\omega}_{xw}^2}}.\end{aligned}$$

In virtù del Teorema dell'indipendenza per trasformazioni descritto ad esempio in Caravenna & Dai Pra (2013, Capitolo 3), si ha che se $X \perp\!\!\!\perp Y^* \mid W$ allora $X \perp\!\!\!\perp Y \mid W$, in quanto Y è una trasformata della variabile Y^* (in modo analogo per $W \perp\!\!\!\perp Y^* \mid X$). Ne consegue che se $W \perp\!\!\!\perp Y \mid X$ allora il parametro $\alpha_w = 0$, di conseguenza la relazione in (3.19) risulta

$$\begin{aligned}\beta(x) = &\frac{\alpha_x}{\sqrt{\omega_{xx}}}\Bigg[\frac{\varphi\big(\tau\sqrt{1+\alpha_x^2}-\frac{\alpha_x}{\sqrt{\omega_{xx}}}\xi_x+\frac{\alpha_x}{\sqrt{\omega_{xx}}}x\big)}{\Phi\big(\tau\sqrt{1+\alpha_x^2}-\frac{\alpha_x}{\sqrt{\omega_{xx}}}\xi_x+\frac{\alpha_x}{\sqrt{\omega_{xx}}}x\big)}+\\ &+\frac{\varphi\big(-\tau\sqrt{1+\alpha_x^2}+\frac{\alpha_x}{\sqrt{\omega_{xx}}}\xi_x-\frac{\alpha_x}{\sqrt{\omega_{xx}}}x\big)}{\Phi\big(-\tau\sqrt{1+\alpha_x^2}+\frac{\alpha_x}{\sqrt{\omega_{xx}}}\xi_x-\frac{\alpha_x}{\sqrt{\omega_{xx}}}x\big)}\Bigg],\end{aligned} \tag{3.20}$$

si nota quindi che la relazione marginale e condizionata in scala di log-rapporto di quote coincidono. Infatti assumendo la nullità del parametro α_w, e derivando rispetto ad x la quantità in definita in (3.10), si ottiene il risultato in (3.20). Tale situazione può essere rappresentata tramite la Figura 3.3(a). Allo stesso modo se $X \perp\!\!\!\perp W$ allora $\bar{\omega}_{xw} = 0$, e si ottiene che $\beta(x)$ risulta

$$\beta(x) = \frac{\alpha_x}{k}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right],$$

tale casistica viene rappresentata in Figura 3.3(b).

In Figura 3.3(c), viene presentato il caso in cui l'effetto di X su Y viene dato solamente dall'effetto mediato, in quanto $X \perp\!\!\!\perp Y \mid W$, si ottiene

$$\begin{aligned}\beta(x) =& \frac{\alpha_w\bar{\omega}_{xw}}{k}\left[\frac{\varphi\big(\tau\sqrt{1+\alpha_w^2} - \frac{\bar{\omega}_{xw}\alpha_w}{k}\xi_w + \frac{\bar{\omega}_{xw}\alpha_w}{k}w\big)}{\Phi\big(\tau\sqrt{1+\alpha_w^2} - \frac{\bar{\omega}_{xw}\alpha_w}{k}\xi_w + \frac{\bar{\omega}_{xw}\alpha_w}{k}w\big)} + \right.\\ &\left. + \frac{\varphi\big(-\tau\sqrt{1+\alpha_w^2} + \frac{\bar{\omega}_{xw}\alpha_w}{k}\xi_w - \frac{\bar{\omega}_{xw}\alpha_w}{k}w\big)}{\Phi\big(-\tau\sqrt{1+\alpha_w^2} + \frac{\bar{\omega}_{xw}\alpha_w}{k}\xi_w - \frac{\bar{\omega}_{xw}\alpha_w}{k}w\big)}\right].\end{aligned}$$

In questo caso la derivata rispetto ad x nel logit condizionato sarebbe pari a 0, mentre non risulta nulla la derivata rispetto ad x nel logit marginale in quanto l'effetto di X su Y risulta pari all'effetto mediato tramite la variabile W.

Capitolo 4

Relazioni lineari nei modelli di regressione per risposta binaria

Nel precedente Capitolo il processo generatore dei dati portava ad avere una forma funzionale non lineare sia nella regressione logistica comprendente i regressori X e W che nella regressione logistica con la sola variabile X. La non linearità porta ad avere dei risultati relativamente complessi in quanto la derivata rispetto al logit marginale è una funzione di x e non un coefficiente costante. Risulta noto che se $(X, W \mid Y = y)$ possiedono distribuzione normale bivariata con matrice di varianze e covarianze uguale per $y = 0$ e per $y = 1$, allora il logit risulta lineare, come mostrato in Anderson (1972). Tale casistica corrisponde alla *LDA (Linear Discriminant Analysis)* introdotta in Fisher (1936), nel caso con $k = 2$ classi. In questo Capitolo si cercherà di ottenere un funzionale lineare a partire da un processo generatore dove la risposta viene definita tramite discretizzazione di una variabile latente Y^*.

4.1 Probit lineare

Utilizzando il processo generatore dati descritto nel Paragrafo 3.3, si è mostrato che la forma funzionale del logit condizionato e del logit marginale risulta non lineare ed inoltre la formula di $\beta(x)$ risulta abbastanza complessa in quanto comprende la somma di *Inverse Mills Ratio*.

L'utilizzo di una *link function* differente potrebbe aiutare nell'ottenere una forma funzionale più semplice ed interpretabile per la regressione binaria. Come descritto ad esempio in Salvan et al. (2020), una funzione legame $g(\cdot)$ per risposta dicotomica deve rispettare le seguenti proprietà:

- $g : [0, 1] \to \mathbb{R}$;
- $g(\cdot)$ monotona crescente.

Ne deriva che appartengono a questa classe tutte le funzioni inverse di funzioni di ripartizione di variabili casuali continue con supporto nei numeri reali.

Utilizzando il *Data Generating Process* definito nella Sezione 3.3 e sfruttando le relazioni definite in (3.7) ed in (3.9), si ottiene

$$\begin{aligned}\mathbb{P}(Y = 1 \mid X = x, W = w) &= \frac{f(x, w \mid Y = 1)\mathbb{P}(Y = 1)}{f(x, w)} \\ &= \frac{\varphi_2(z - \xi; \Omega)\Phi(\alpha_0 + \alpha^\top \omega^{-1}(z - \xi))\Phi(\tau)^{-1}\Phi(\tau)}{\varphi_2(z - \xi; \Omega)} \\ &= \Phi(\alpha_0 + \alpha^\top \omega^{-1}(z - \xi)),\end{aligned}$$

dove $z = (x, w)$, $\alpha = (\alpha_x, \alpha_w)$ e ω è una matrice diagonale 2×2 contente gli elementi $(\sqrt{\omega_{xx}}, \sqrt{\omega_{ww}})$. A questo punto viene naturale utilizzare la funzione *probit* come funzione legame al fine di avere una forma funzionale lineare rispetto alle covariate di interesse, infatti

$$\begin{aligned}\Phi^{-1}\{\mathbb{P}(Y = 1 \mid X = x, W = w)\} &= \alpha_0 + \alpha^\top \omega^{-1}(z - \xi) \\ &= \alpha_0 - \alpha^\top \omega^{-1}\xi + \alpha^\top \omega^{-1} z \\ &= \alpha_0 - \alpha^\top \omega^{-1}\xi + \frac{\alpha_x}{\sqrt{\omega_{xx}}}x + \frac{\alpha_w}{\sqrt{\omega_{ww}}}w,\end{aligned}$$

ne deriva che il parametro condizionato rispetto alla variabile x risulta pari a $\frac{\alpha_x}{\sqrt{\omega_{xx}}}$.

In modo analogo si può procedere per il calcolo del probit marginale rispetto ad x. Come dimostrato in (3.12), la distribuzione marginale di X dato Y rimane

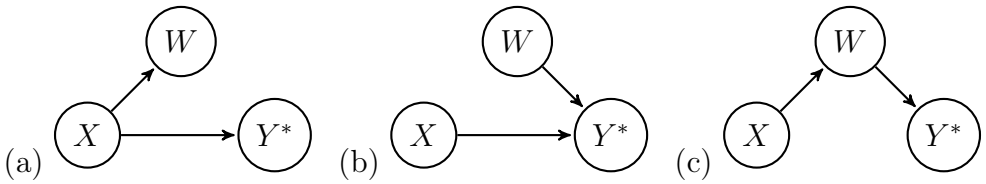

Figura 4.1: Grafico delle relazioni (a) $W \perp\!\!\!\perp Y^* \mid X$ e (b) $W \perp\!\!\!\perp X$ e (c) $X \perp\!\!\!\perp Y^* \mid W$.

Extended Skew Normal con medesimo parametro τ. Si può quindi definire

$$\begin{aligned}\Phi^{-1}\{\mathbb{P}(Y=1 \mid X=x)\} &= \Phi^{-1}\left\{\Phi(\tau\sqrt{1+\alpha^2_{x(w)}} + \alpha_{x(w)}\omega_{xx}^{-\frac{1}{2}}(x-\xi_x))\right\} \\ &= \tau\sqrt{1+\alpha^2_{x(w)}} - \frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}\xi_x + \frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}}x\end{aligned} \tag{4.1}$$

Come si può vedere, la linearità viene mantenuta anche nel caso marginale e $\beta(x)$ risulta quindi costante e pari a

$$\frac{\alpha_{x(w)}}{\sqrt{\omega_{xx}}} = \frac{\alpha_x + \bar{\omega}_{xw}\alpha_w}{\sqrt{1+\alpha_w^2(1-\bar{\omega}^2_{xw})}}\frac{1}{\sqrt{\omega_{xx}}}.$$

Un risultato simile in termini di *path analysis* viene ottenuto in Winship & Mare (1983) a partire da modelli lineari dicotomizzati e non da una distribuzione multivariata come nel caso in esame. Anche in questo caso risulta possibile trarre delle conclusioni in termini di relazioni marginali tra le variabili di interesse.

Come descritto nel Paragrafo 3.5.1, se $W \perp\!\!\!\perp Y \mid X$ allora il parametro α_w risulta nullo, e si ottiene che il parametro marginale risulta uguale al parametro condizionato. In relazione con quanto scritto nel Paragrafo 3.1, se $W \perp\!\!\!\perp X$ allora $\bar{\omega}_{xw} = 0$ e ne deriva che

$$\beta(x) = \frac{\alpha_x}{\sqrt{\omega_{xx}}}\frac{1}{\sqrt{1+\alpha_w^2}},$$

essendo la quantità $\frac{1}{\sqrt{1+\alpha_w^2}}$ compresa nell'intervallo tra 0 e 1, si ottiene che il parametro marginale è in valore assoluto minore del parametro condizionato, in linea con il risultato di Neuhaus & Jewell (1993). Se invece il parametro $\alpha_x = 0$ si ricade nel caso in cui $X \perp\!\!\!\perp Y \mid W$, ne consegue che

$$\beta(x) = \frac{\bar{\omega}_{xw}\alpha_w}{\sqrt{1+\alpha_w^2(1-\bar{\omega}^2_{xw})}}\frac{1}{\sqrt{\omega_{xx}}}.$$

In questo circostanza l'effetto della variabile X sulla variabile Y viene dato dall'effetto indiretto mediato tramite la variabile W. Le 3 casistiche sopra definite sono descritte nei grafici in Figura 4.1.

4.2 Descrizione di un nuovo *Data Generating Process*

Riprendendo le problematiche discusse nei Paragrafi 1.3 e 1.4, si vuole definire un meccanismo generatore dei dati tale per cui l'assunzione di linearità sia rispettata nel logit condizionato. A partire da tale assunzione si cercherà di capire in quali casi sia realistico postulare un logit marginale lineare.

Sia $Z = (X, W) \sim N_2(0, \bar{\Omega})$ il vettore delle covariate di interesse utili nello spiegare la risposta, dove il coefficiente $\bar{\omega}_{xw}$ presente in $\bar{\Omega}$ indica la correlazione tra X e W. Si definisca la variabile latente

$$Y^* = \alpha^\top Z - T = \alpha_x X + \alpha_w W - T$$

con $T \sim Lo(0, 1)$ indipendente da Z, dove T può essere interpretato come errore additivo alla combinazione lineare tra X e W. La variabile aleatoria Y^* ha densità non nota ma risulta comunque possibile calcolarne alcuni momenti

$$\begin{aligned} \mathbb{E}[Y^*] &= 0, \\ \mathbb{V}ar[Y^*] &= \mathbb{E}[(Y^*)^2] = \alpha^\top \bar{\Omega} \alpha + \frac{\pi^2}{3}, \\ \mathbb{E}[(Y^*)^3] &= 0, \end{aligned}$$

in particolare essendo il momento terzo nullo allora la variabile casuale risulta simmetrica e quindi $\mathbb{P}(Y^* > 0) = \frac{1}{2}$. La funzione generatrice dei momenti per la variabile Y^* è pari a

$$M_{Y^*}(t) = \frac{M_Z(t)}{M_T(t)} = \frac{\exp(\frac{s^2 t^2}{2})}{B(1-t, 1+t)}, \quad \text{per } t \in (-1, 1),$$

dove $s^2 = \alpha^\top \bar{\Omega} \alpha$ mentre $B(\cdot, \cdot)$ indica la funzione Beta. Essendo la funzione generatrice dei momenti finita in un intervallo contente lo zero allora la variabile casuale presenta momenti finiti di ogni ordine. Risulta inoltre possibile calcolare il vettore

bidimensionale contenente la correlazione tra $Z = (X, W)$ e Y^*, in particolare

$$\text{corr}(Z, Y^*) = \delta = \left(\frac{\pi^2}{3} + \alpha^\top \bar{\Omega}\alpha\right)^{-\frac{1}{2}} \bar{\Omega}\alpha.$$

La variabile risposta viene definita tramite la seguente discretizzazione della variabile latente Y^*,

$$Y = \begin{cases} 1 & \text{se } Y^* > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

ne deriva che $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$.

Utilizzando il Lemma 2 e l'equazione definita in (2.1), si ottiene che la densità della variabile aleatoria $(Z \mid Y = 1) = (X, W \mid Y = 1)$ risulta

$$f(z \mid Y = 1) = 2\varphi_2(z; \bar{\Omega}) \frac{\exp(\alpha^\top z)}{1 + \exp(\alpha^\top z)}, \tag{4.2}$$

e verrà indicata da qui in avanti tramite la sigla $SL(0, \bar{\Omega}, \alpha)$. Allo stesso modo si ottiene che la variabile casuale $(Z \mid Y = 0) = (X, W \mid Y = 0)$ presenta distribuzione $SL(0, \bar{\Omega}, -\alpha)$.

Definiti tali elementi risulta possibile calcolare $\mathbb{P}(Y = 1 \mid X = x, W = w)$; utilizzando il teorema di Bayes infatti si ha

$$\begin{aligned} \mathbb{P}(Y = 1 \mid X = x, W = w) &= \frac{f(x, w \mid Y = 1)\mathbb{P}(Y = 1)}{f(x, w)} \\ &= 2\varphi_2(z, \bar{\Omega}) \frac{\exp(\alpha^\top z)}{1 + \exp(\alpha^\top z)} \frac{1}{2} \varphi_2(z, \bar{\Omega})^{-1} \\ &= \frac{\exp(\alpha^\top z)}{1 + \exp(\alpha^\top z)}. \end{aligned}$$

Utilizzando la funzione legame logistica per la regressione per dati binari, si ottiene

$$\log \frac{\mathbb{P}(Y = 1 | X = x, W = w)}{\mathbb{P}(Y = 0 | X = x, W = w)} = \alpha_x x + \alpha_w w \tag{4.3}$$

cioè una funzione lineare rispetto alle covariate di interesse, in linea con il risultato

mostrato in MacKinnon et al. (2007). Risulta infatti equivalente scrivere

$$\begin{aligned}\mathbb{P}(Y=1 \mid X=x, W=w) =& \mathbb{P}(\alpha_x x + \alpha_w w - T > 0) \\ =& \mathbb{P}(T < \alpha_x x + \alpha_w w) \\ =& \frac{\exp(\alpha_x x + \alpha_w w)}{1+\exp(\alpha_x x + \alpha_w w)},\end{aligned}$$

in quanto T presenta distribuzione logistica standard.

4.2.1 Relazione marginale

Per utilizzare la formula in equazione (3.11) e ricavare la relazione di X rispetto ad Y, risulta necessario esplicitare la distribuzione marginale di X condizionatamente al valore assunto dalla risposta.

Integrando la densità congiunta rispetto a w, si ha che la densità marginale di X dato il valore assunto da Y risulta

$$\begin{aligned}f(x \mid Y=1) =& \int_{-\infty}^{+\infty} f(x, w \mid Y=1)\, dw = \int_{-\infty}^{+\infty} 2\varphi_2(z; \bar{\Omega}) \frac{\exp(\alpha_x x + \alpha_w w)}{1+\exp(\alpha_x x + \alpha_w w)}\, dw \\ =& 2\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi(1-\bar{\omega}_{xw}^2)}} \exp\left\{ -\frac{1}{2}\left(\frac{w^2 - 2\bar{\omega}_{xw} w x + \bar{\omega}_{xw}^2 x^2}{1-\bar{\omega}_{xw}^2} \right) \right\} \cdot \\ & \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x^2 - \bar{\omega}_{xw}^2 x^2}{1-\bar{\omega}_{xw}^2} \right) \right\} \frac{\exp(\alpha_x x + \alpha_w w)}{1+\exp(\alpha_x x + \alpha_w w)}\, dw \\ =& \frac{2}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi(1-\bar{\omega}_{xw}^2)}} \exp\left\{ -\frac{1}{2}\left(\frac{w - \bar{\omega}_{xw} x}{\sqrt{1-\bar{\omega}_{xw}^2}} \right)^2 \right\} \cdot \\ & \cdot \frac{\exp(\alpha_x x + \alpha_w w)}{1+\exp(\alpha_x x + \alpha_w w)}\, dw,\end{aligned}$$

effettuando la sostituzione $s = \frac{w - \bar{\omega}_{xw} x}{\sqrt{1-\bar{\omega}_{xw}^2}}$,

$$\begin{aligned}f(x \mid Y=1) =& \frac{2}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} s^2 \right\} \cdot \\ & \cdot \frac{\exp((\alpha_x + \alpha_w \bar{\omega}_{xw}) x + \alpha_w \sqrt{1-\bar{\omega}_{xw}^2}\, s)}{1+\exp((\alpha_x + \alpha_w \bar{\omega}_{xw}) x + \alpha_w \sqrt{1-\bar{\omega}_{xw}^2}\, s)}\, ds,\end{aligned}$$

Si noti che all'interno dell'integrale appare il nucleo di una densità asimmetri-

ca estesa definita in equazione (2.16), la cui costante di normalizzazione risulta $\mathbb{P}\{\alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* - T > -(\alpha_x+\alpha_w\bar{\omega}_{xw})x\}$, dove T risulta essere una $Lo(0,1)$ mentre Z_0^* indica una variabile unidimensionale con distribuzione normale standard. La densità risulta quindi

$$f(x \mid Y=1) = 2\,\varphi(x)\,\mathbb{P}\big\{\alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* - T > -(\alpha_x+\alpha_w\bar{\omega}_{xw})x\big\},$$

che rimane SL solamente se il parametro α_w risulta pari a 0, in caso contrario la distribuzione non ha forma esplicita in quanto non è nota la funzione di ripartizione di una variabile casuale formata dalla differenza tra una variabile normale ed una logistica standard. In maniera analoga, la densità della variabile X dato $Y=0$ risulta

$$f(x \mid Y=0) = 2\,\varphi(x)\,\mathbb{P}\big\{-\alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* - T > (\alpha_x+\alpha_w\bar{\omega}_{xw})x\big\}.$$

Indicando con V la variabile aleatoria $V = \alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* - T$, si ha che V ha densità simmetrica e $\mathbb{V}ar[V] = \alpha_w^2(1-\bar{\omega}_{xw}^2) + \frac{\pi^2}{3}$. Si vuole quindi approssimare la variabile V attraverso una distribuzione logistica che presenta la stessa varianza di V al fine di ottenere una forma marginale per X che risulti approssimativamente SL. Si definisca quindi la variabile

$$V^a \sim Lo\left(0, \sqrt{\frac{3}{\pi^2}\left(\frac{\pi^2}{3} + \alpha_w^2(1-\bar{\omega}_{xw}^2)\right)}\right),$$

dove i momenti risultano

$$\begin{aligned}
\mathbb{E}[V^a] &= 0,\\
\mathbb{V}ar[V^a] &= \alpha_w^2(1-\bar{\omega}_{xw}^2) + \frac{\pi^2}{3},\\
\mathbb{E}[(V^a)^3] &= 0.
\end{aligned}$$

Essendo la logistica una famiglia di scala, risulta possibile approssimare $\mathbb{P}\{\alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* -$

$T > -(\alpha_x + \alpha_w\bar{\omega}_{xw})x\}$ tramite

$$\begin{aligned}\mathbb{P}\{\alpha_w\sqrt{1-\bar{\omega}_{xw}^2}Z_0^* - T > -(\alpha_x + \alpha_w\bar{\omega}_{xw})x\} &\approx \mathbb{P}\{V^a > -(\alpha_x + \alpha_w\bar{\omega}_{xw})x\} \\ &= 1 - \mathbb{P}\{V^a < -(\alpha_x + \alpha_w\bar{\omega}_{xw})x\} \\ &= \frac{\exp\left(\frac{\pi}{\sqrt{3}}\frac{(\alpha_x+\alpha_w\bar{\omega}_{xw})}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2)+\frac{\pi^2}{3}}}x\right)}{1+\exp\left(\frac{\pi}{\sqrt{3}}\frac{(\alpha_x+\alpha_w\bar{\omega}_{xw})}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2)+\frac{\pi^2}{3}}}x\right)}\end{aligned}$$

Utilizzando questa approssimazione risulta possibile scrivere

$$\begin{aligned}(X \mid Y = 1) &\overset{appr}{\sim} SL(0, 1, \alpha_{x(w)}), \\ (X \mid Y = 0) &\overset{appr}{\sim} SL(0, 1, -\alpha_{x(w)}),\end{aligned} \tag{4.4}$$

dove il parametro $\alpha_{x(w)}$ viene definito tramite

$$\alpha_{x(w)} = \frac{\pi}{\sqrt{3}}\frac{(\alpha_x + \alpha_w\bar{\omega}_{xw})}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2) + \frac{\pi^2}{3}}}.$$

A questo punto risulta possibile derivare il parametro marginale utilizzando la formula (3.11), ottenendo

$$\begin{aligned}\log\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} &= \log\frac{f(x \mid Y=1)}{f(x \mid Y=0)} + \log\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} \\ &\approx \alpha_{x(w)}x,\end{aligned} \tag{4.5}$$

dove l'approssimazione deriva dal fatto che la distribuzione marginale non è esatta ma approssimata. Ne consegue che il parametro marginale risulta costante e pari ad $\alpha_{x(w)}$.

4.3 Studio di simulazione

L'approssimazione per la distribuzione marginale definita in (4.4), si basa essenzialmente sull'uguaglianza del momento secondo di una distribuzione logistica e di una variabile casuale derivante dalla somma tra una normale e una logistica standard. Tale approssimazione risulta utile al fine di definire come l'effetto totale

si distribuisce tra effetto diretto ed indiretto, nonostante ciò risulta di interesse capire l'errore che si commette nel postulare una funzione lineare anche per il modello marginale.

Al fine di valutare la bontà dell'approssimazione, si è proceduto con uno studio di simulazione che a partire dal processo generatore definito nel Paragrafo 4.2, va a stimare i parametri della regressione logistica utilizzando il metodo della massima verosimiglianza. Vengono successivamente effettuati dei *test alla Wald* per verificare le ipotesi sui parametri del modello condizionato e del modello marginale.

Come si nota dalle Tabelle riportate in Appendice B, per il modello condizionato si nota che per ogni scelta di α e di n, viene accettata l'ipotesi nulla circa il 95% delle volte e i $\hat{\beta}$ corrispondono all'incirca ai valori teorici, infatti dalla teoria si ha $\beta_x = 0, \beta_x = \alpha_x, \beta_w = \alpha_w$. Al contrario per il modello marginale si ha che per valori piccoli di α_w l'ipotesi sui singoli parametri viene accettata all'incirca il 95% delle volte, mentre al crescere di α_w e di n la frazione di volte per cui si accettano le ipotesi $H_0 : \beta_x = \alpha_{x(w)}$ e $H_0 : \beta_0 = 0, \beta_x = \alpha_{x(w)}$ scende drasticamente.

4.4 Alcuni commenti

Dallo studio di simulazione si nota come l'approssimazione presentata in (4.5), sia da ritenersi soddisfacente solamente per valori di α_w vicini allo 0 in valore assoluto. Risulta di interesse notare che il processo generatore dei dati descritto nel Paragrafo 4.2 postula un perfetto bilanciamento tra le due classi in quanto $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$. L'aggiunta di un parametro α_0 come fatto in equazione (2.16), nonostante conduca a delle complicazioni a livello matematico, porta a risultati molto simili rispetto al caso precedente. Postulando per i dati il processo generatore descritto nel Paragrafo 4.2, ma assumendo che la variabile risposta Y sia definita dalla seguente discretizzazione

$$Y = \begin{cases} 1 & \text{se } Y^* > -\alpha_0, \\ 0 & \text{altrimenti,} \end{cases}$$

si ottiene che $\mathbb{P}(Y = 1) = \mathbb{P}(Y^* > -\alpha_0) = \mathbb{P}(\alpha^\top Z - T > -\alpha_0)$. Sfruttando il risultato in (2.16), la distribuzione delle covariate rispetto al valore della risposta

risulta avere densità rispettivamente pari a

$$\begin{aligned} f(x, w \mid Y = 1) &= \varphi_2(z; \bar{\Omega}) \frac{\exp(\alpha_0 + \alpha^\top z)}{1 + \exp(\alpha_0 + \alpha^\top z)} \mathbb{P}(\alpha^\top Z - T > -\alpha_0)^{-1}, \\ f(x, w \mid Y = 0) &= \varphi_2(z; \bar{\Omega}) \frac{\exp(-\alpha_0 - \alpha^\top z)}{1 + \exp(-\alpha_0 - \alpha^\top z)} \mathbb{P}(-\alpha^\top Z - T > \alpha_0)^{-1}, \end{aligned} \tag{4.6}$$

in relazione al risultato presentato nel Paragrafo 2.5.

Utilizzando la relazione (3.9), si ha che il logit condizionato risulta

$$\log \frac{\mathbb{P}(Y = 1 \mid X = x, W = w)}{\mathbb{P}(Y = 0 \mid X = x, W = w)} = \alpha_0 + \alpha_x x + \alpha_w w. \tag{4.7}$$

Per ottenere il logit marginale risulta necessario ricavare la distribuzione marginale di X condizionatamente al valore di Y; in particolare integrando le densità in (4.6) rispetto a w, si ottiene

$$\begin{aligned} f(x \mid Y = 1) = &\int_{-\infty}^{+\infty} f(x, w \mid Y = 1)\, dw \\ = &\frac{\mathbb{P}(\alpha^\top Z - T > -\alpha_0)^{-1}}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi(1 - \bar{\omega}_{xw}^2)}} \cdot \\ &\cdot \exp\left\{ -\frac{1}{2(1 - \bar{\omega}_{xw}^2)} (w - \bar{\omega}_{xw} x)^2 \right\} \frac{\exp(\alpha_0 + \alpha_x x + \alpha_w w)}{1 + \exp(\alpha_0 + \alpha_x x + \alpha_w w)}\, dw, \end{aligned}$$

effettuando la sostituzione $s = \frac{w - \bar{\omega}_{xw} x}{\sqrt{1 - \bar{\omega}_{xw}^2}}$,

$$\begin{aligned} f(x \mid Y = 1) = &\frac{\mathbb{P}(\alpha^\top Z - T > -\alpha_0)^{-1}}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} s^2 \right\} \cdot \\ &\cdot \frac{\exp(\alpha_0 + (\alpha_x + \alpha_w \bar{\omega}_{xw}) x + \alpha_w \sqrt{1 - \bar{\omega}_{xw}^2} s)}{1 + \exp(\alpha_0 + (\alpha_x + \alpha_w \bar{\omega}_{xw}) x + \alpha_w \sqrt{1 - \bar{\omega}_{xw}^2} s)}\, ds \\ = &\frac{\varphi(x)}{\mathbb{P}(\alpha^\top Z - T > -\alpha_0)} \mathbb{P}(\alpha_w \sqrt{1 - \bar{\omega}_{xw}^2} Z_0^* - T > -\alpha_0 - (\alpha_x + \alpha_w \bar{\omega}_{xw}) x), \end{aligned}$$

con $Z_0^* \sim N(0, 1)$. Allo stesso modo la densità marginale di X rispetto ad $Y = 0$ risulta

$$f(x \mid Y = 0) = \frac{\varphi(x)}{\mathbb{P}(\alpha^\top Z - T > \alpha_0)} \mathbb{P}(-\alpha_w \sqrt{1 - \bar{\omega}_{xw}^2} Z_0^* - T > \alpha_0 + (\alpha_x + \alpha_w \bar{\omega}_{xw}) x)$$

Utilizzando la formula in (3.11), risulta

$$\begin{aligned}\log\frac{\mathbb{P}(Y=1\mid X=x)}{\mathbb{P}(Y=0\mid X=x)} &= \log\frac{f(x\mid Y=1)}{f(x\mid Y=0)} + \log\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} \\ &= \log\frac{\varphi(x)\mathbb{P}(\alpha^\top Z - T > -\alpha_0)^{-1}}{\varphi(x)\mathbb{P}(\alpha^\top Z - T > \alpha_0)^{-1}} + \log\frac{\mathbb{P}(\alpha^\top Z - T > -\alpha_0)}{\mathbb{P}(\alpha^\top Z - T > \alpha_0)} + \\ &\quad + \log\frac{\mathbb{P}(\alpha_w\sqrt{1-\bar\omega_{xw}^2}Z_0^* - T > -\alpha_0 - (\alpha_x + \alpha_w\bar\omega_{xw})x)}{\mathbb{P}(-\alpha_w\sqrt{1-\bar\omega_{xw}^2}Z_0^* - T > \alpha_0 + (\alpha_x + \alpha_w\bar\omega_{xw})x)} \\ &= \log\frac{\mathbb{P}(\alpha_w\sqrt{1-\bar\omega_{xw}}Z_0^* - T > -\alpha_0 - (\alpha_x + \alpha_w\bar\omega_{xw})x)}{\mathbb{P}(-\alpha_w\sqrt{1-\bar\omega_{xw}^2}Z_0^* - T > \alpha_0 + (\alpha_x + \alpha_w\bar\omega_{xw})x)},\end{aligned}$$

A questo punto, utilizzando le approssimazioni precedentemente definite, si può approssimare la quantità a destra del segno di uguaglianza tramite

$$\begin{aligned}\log\frac{\mathbb{P}(Y=1\mid X=x)}{\mathbb{P}(Y=0\mid X=x)} &\approx \log\frac{\exp\left(\frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}} + \frac{\pi}{\sqrt{3}}\frac{\alpha_x+\alpha_w\bar\omega_{xw}}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}}x\right)}{1+\exp\left(\frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}} + \frac{\pi}{\sqrt{3}}\frac{\alpha_x+\alpha_w\bar\omega_{xw}}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}}x\right)} + \\ &\quad - \log\frac{\exp\left(-\frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}} - \frac{\pi}{\sqrt{3}}\frac{\alpha_x+\alpha_w\bar\omega_{xw}}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}}x\right)}{1+\exp\left(-\frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}} - \frac{\pi}{\sqrt{3}}\frac{\alpha_x+\alpha_w\bar\omega_{xw}}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}}x\right)} \\ &= \frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}} + \frac{\pi}{\sqrt{3}}\frac{\alpha_x+\alpha_w\bar\omega_{xw}}{\sqrt{\alpha_w^2(1-\bar\omega_{xw}^2)+\frac{\pi^2}{3}}}x,\end{aligned} \tag{4.8}$$

che risulta lineare rispetto ad x. Il parametro marginale risulta quindi costante e pari ad $\alpha_{x(w)}$.

4.4.1 Suddivisione dell'effetto totale

A questo punto viene naturale vedere come l'effetto totale va a suddividersi tra effetto diretto ed indiretto. Richiamando i risultati in (4.1), vista la normalità congiunta delle covariate, si ottiene che le variabili X e W risultano indipendenti solamente se il parametro $\bar\omega_{xw}$ risulta nullo. In questo caso, il parametro marginale

risulta circa pari a

$$\frac{\pi}{\sqrt{3}} \frac{\alpha_x}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2)+\frac{\pi^2}{3}}}, \tag{4.9}$$

ed essendo la quantità $\frac{\pi/\sqrt{3}}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2)+\frac{\pi^2}{3}}} \in (0,1]$, si ottiene che il parametro marginale risulta in valore assoluto minore rispetto al parametro condizionato. Nel caso in cui $\alpha_w = 0$, si ha che $W \perp\!\!\!\perp Y \mid X$, ne deriva che il parametro marginale risulta esattamente pari ad α_x, di conseguenza si ottiene che il parametro marginale e il parametro condizionato coincidono. Nel caso in cui $X \perp\!\!\!\perp Y \mid W$, si ha che $\alpha_x = 0$ e il parametro marginale risulta

$$\frac{\pi}{\sqrt{3}} \frac{\alpha_w \bar{\omega}_{xw}}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2)+\frac{\pi^2}{3}}}.$$

Tali relazioni, come anticipato, valgono solamente approssimativamente per valori piccoli in valore assoluto di α_w.

Capitolo 5

Applicazione a dati reali

In questo Capitolo viene presentata l'applicazione delle procedure definite in precedenza ad un insieme di dati reali con il fine di evidenziare potenzialità e punti deboli delle metodologie presentate.

I dati scelti sono presenti nel lavoro di Cortez et al. (2009) e riguardano le caratteristiche del vino portoghese *vinho verde*: una rinomata tipologia di vino prodotta nel nord-ovest del paese. La variabile risposta è rappresentata dalla qualità del vino espressa tramite una votazione che va da 0 a 10, ottenuta valutando ogni campione attraverso tre valutatori sensoriali. Al fine di ottenere una risposta binaria si è effettuata la seguente dicotomizzazione:

- se il voto risulta maggiore o uguale a 6 si ha un vino di **buona** qualità (indicato con la classe 1);
- se il voto risulta minore di 6 si ha un vino di **scarsa** qualità (indicato con la classe 0).

I regressori sono rappresentati da caratteristiche fisico-chimico del vino, come ad esempio il pH, la densità, i solfiti o l'acido citrico. In particolare, si è scelto di utilizzare come covariata d'interesse l'acidità volatile, mentre come mediatore si è utilizzato il logaritmo dei solfiti misurati in mg/l. Risulta infatti noto che un livello troppo elevato di acidità volatile può portare il vino ad avere uno sgradevole sapore di aceto, tuttavia tale caratteristica deve essere valutata in relazione con altre qualità fisico-chimiche come i solfiti, il pH o gli zuccheri residui.

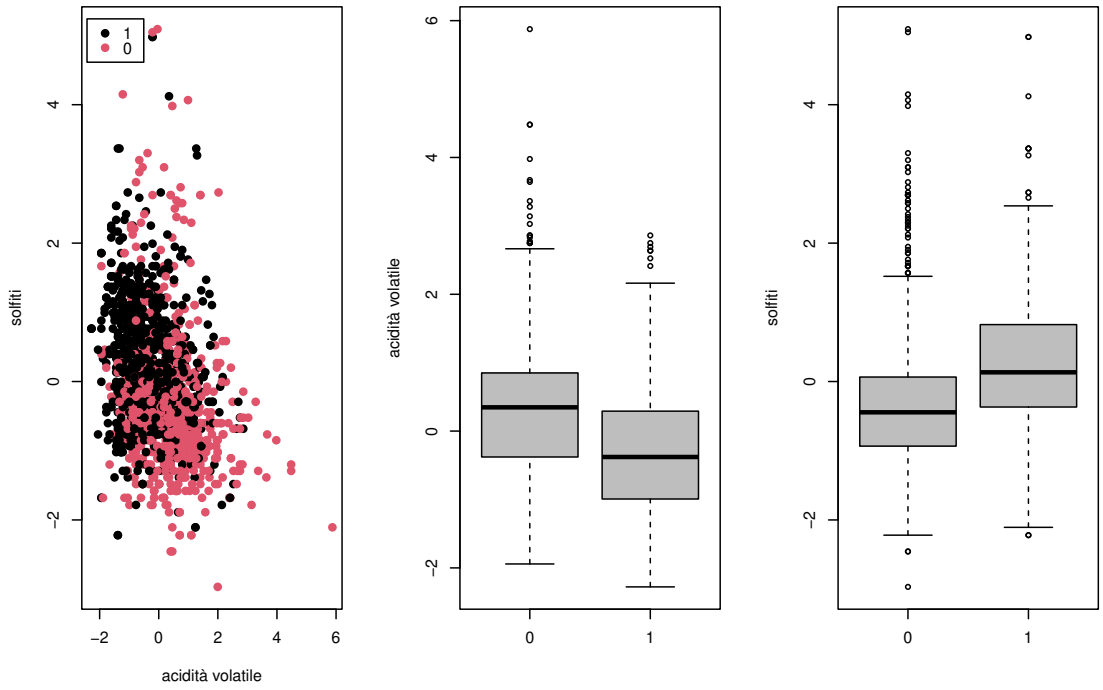

Figura 5.1: Scatter plot delle covariate rispetto al valore della risposta (in nero la distribuzione rispetto ad $y = 1$, in rosso rispetto ad $y = 0$) e boxplot delle covariate rispetto alla variabile risposta.

Il dataset contiene un totale di 1599 osservazioni, di cui 855 presentano una qualità del vino buona mentre le restanti 744 presentano una qualità ritenuta non sufficiente. Al fine di applicare le metodologie presentate nei precedenti Capitoli si è effettuata una standardizzazione delle covariate. Dai grafici in Figura 5.1 si nota come per valori più elevati della variabile acidità volatile si ha una qualità del vino più bassa, mentre un livello più elevato di solfiti porta ad avere vini più gradevoli.

5.1 Applicazione delle metodologie

A questo punto risulta d'interesse valutare come si suddivide l'effetto totale di X (acidità volatile) su Y in effetti diretti ed in effetti mediati tramite la variabile W (solfiti in scala logaritmica). Postulando il processo generatore descritto nel

Paragrafo 3.3 e utilizzando la metodologia presentata nel Paragrafo 4.1, si ha

$$g\{\mathbb{P}(Y = 1 \mid X = x, W = w)\} = \beta_0 + \beta_x x + \beta_w w$$

con $g(\cdot) = \Phi(\cdot)$. Essendo la deviazione standard delle covariate pari a 1, si ha che $\beta_x = \frac{\alpha_x}{\sqrt{\omega_{xx}}} = \alpha_x$ e $\beta_w = \frac{\alpha_w}{\sqrt{\omega_{ww}}} = \alpha_w$. Il modello marginale rimane lineare, in particolare

$$g\{\mathbb{P}(Y = 1 \mid X = x)\} = \eta_0 + \eta_x x,$$

dove $g(\cdot) = \Phi(\cdot)$. In particolare, si è mostrato in equazione (4.1) che il processo generatore porta ad avere

$$\eta_x = \frac{\alpha_x + \bar{\omega}_{xw}\alpha_w}{\sqrt{1 + \alpha_w^2(1 - \bar{\omega}_{xw}^2)}} \frac{1}{\sqrt{\omega_{xx}}}.$$

Utilizzando il metodo della massima verosimiglianza per stimare i parametri, si ottengono i seguenti risultati

$$\hat{\beta}_x = \hat{\alpha}_x = -0.3809 \quad \hat{\beta}_w = \hat{\alpha}_w = 0.2621, \tag{5.1}$$

$$\hat{\eta}_x = -0.4466, \tag{5.2}$$

mentre la stima per la correlazione tra X e W risulta essere $\hat{\omega}_{xw} = -0.3005$. La differenza tra i parametri $\hat{\eta}_x$ e $\hat{\beta}_x$ risulta essere pari a -0.066; tale stima, come dimostrato nei precedenti Capitoli, non quantifica l'effetto indiretto di X sulla risposta. Una stima per gli effetti indiretti è data da

$$IE = \frac{\hat{\omega}_{xw}\hat{\alpha}_w}{\sqrt{1 + \hat{\alpha}_w^2(1 - \hat{\omega}_{xw})}} = -0.076,$$

che risulta maggiore in valore assoluto rispetto alla semplice differenza tra i due coefficienti. Utilizzando la funzione legame logistica per il meccanismo generatore dati definito al Paragrafo 3.3, si ottiene

$$\log \frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = 0 \mid X = x)} = \log \frac{\Phi\left(\alpha_0 - \alpha^\top \omega^{-1}\xi + \frac{\alpha_x}{\sqrt{\omega_{xx}}}x + \frac{\alpha_w}{\sqrt{\omega_{ww}}}w\right)}{1 - \Phi\left(\alpha_0 - \alpha^\top \omega^{-1}\xi + \frac{\alpha_x}{\sqrt{\omega_{xx}}}x + \frac{\alpha_w}{\sqrt{\omega_{ww}}}w\right)}$$

e la relazione tra il parametro marginale e il parametro condizionato risulta essere non lineare, e pari a

$$\beta(x) = \frac{\alpha_x}{k}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right] + \frac{\alpha_w\bar{\omega}_{xw}}{k}\left[\frac{\varphi(\tau_{w.x}(x))}{\Phi(\tau_{w.x}(x))} + \frac{\varphi(-\tau_{w.x}(x))}{\Phi(-\tau_{w.x}(x))}\right],$$

con $k = \sqrt{1 + \alpha_w^2(1 - \bar{\omega}_{xw}^2)}$ e dove $\tau_{w.x}$ viene definito in equazione (3.15). Stimando tramite massima verosimiglianza si ottengono risultati praticamente coincidenti con l'equazione (5.1). Una stima dell'effetto indiretto in scala logit, come definito in equazione (3.19), è data da a

$$IE(x) = \frac{\hat{\alpha}_w\hat{\omega}_{xw}}{k}\left[\frac{\varphi(\hat{\tau}_{w.x}(x))}{\Phi(\hat{\tau}_{w.x}(x))} + \frac{\varphi(-\hat{\tau}_{w.x}(x))}{\Phi(-\hat{\tau}_{w.x}(x))}\right],$$

la notazione $IE(x)$ è dovuta al fatto che l'effetto non è costante in x.

Postulando il processo generatore definito in 4.2, si ottiene un logit lineare per il modello condizionato

$$\begin{aligned}\log\frac{\mathbb{P}(Y=1 \mid X=x, W=w)}{\mathbb{P}(Y=0 \mid X=x, W=w)} &= \beta_0 + \beta_x x + \beta_w w \\ &= \alpha_0 + \alpha_x x + \alpha_w w,\end{aligned}$$

mentre per il modello marginale, come mostrato in (4.8), si ottiene

$$\log\frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} \approx \frac{\pi}{\sqrt{3}}\frac{\alpha_0}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2) + \frac{\pi^2}{3}}} + \alpha_{x(w)}x$$

dove il parametro $\alpha_{x(w)}$ risulta essere pari a

$$\alpha_{x(w)} = \frac{\pi}{\sqrt{3}}\frac{\alpha_x + \alpha_w\bar{\omega}_{xw}}{\sqrt{\alpha_w^2(1-\bar{\omega}_{xw}^2) + \frac{\pi^2}{3}}}.$$

La stima dei parametri ottenuta attraverso il metodo della massima verosimi-

glianza, in questo caso risulta essere pari a

$$\hat{\beta}_x = \hat{\alpha}_x = -0.6207 \quad \hat{\beta}_w = \hat{\alpha}_w = 0.4572,$$
$$\hat{\eta}_x = -0.7300.$$

Anche in questo la differenza tra i coefficienti, che risulta essere $\hat{\eta}_x - \hat{\alpha}_x = -0.1092$, sottostima l'effetto indiretto ottenuto tramite la formula

$$IE \approx \frac{\pi}{\sqrt{3}} \frac{\hat{\alpha}_w \hat{\omega}_{xw}}{\sqrt{\hat{\alpha}_w^2(1-\hat{\omega}_{xw}^2) + \frac{\pi^2}{3}}} = -0.130.$$

Una stima per l'effetto diretto è data da

$$DE \approx \frac{\pi}{\sqrt{3}} \frac{\hat{\alpha}_x}{\sqrt{\hat{\alpha}_w^2(1-\hat{\omega}_{xw}^2) + \frac{\pi^2}{3}}} = -0.570,$$

che risulta diversa da $\hat{\beta}_x$, come visto in precedenza. La stima di β_x porterebbe quindi a sovrastimare gli effetti diretti in valore assoluto.

Conclusioni

In questo lavoro si è cercato di approfondire il tema della *mediation analysis* applicata a modelli per risposta binaria. Questa tematica risulta abbastanza complessa nel caso in cui la risposta abbia natura dicotomica in quanto le relazioni tra parametri marginali e condizionati si fanno complicate; inoltre le proprietà di scomposizione dell'effetto totale valide nel contesto dei modelli lineari per risposta e mediatori continui non valgono se applicate in questa situazione. Le principali problematiche derivano dalla non collassabilità delle funzioni legame (Neuhaus & Jewell, 1993) e dalla forma funzionale assunta dal modello marginale (Greenland et al., 1999).

In particolare, nel Capitolo 3, partendo da un processo generatore definito sulla base di una variabile normale multivariata dove una delle variabili rappresenta una versione latente della risposta, è stata trovata la relazione che lega parametri condizionati e marginali nel caso di regressori continui. Tale formula permette inoltre la divisione dell'effetto totale in effetti diretti e indiretti. Le principali problematiche sono legate alla non linearità del logit sia nel modello condizionato che nel modello marginale.

Nel Capitolo 4, si è cercato di risolvere il problema di ottenere una forma lineare nel modello marginale. In prima istanza, a partire dal processo generatore definito in precedenza, si è utilizzata una funzione legame che permettesse di avere un predittore lineare rispetto alle covariate nel modello condizionato e nel modello ridotto. Successivamente si è trovata la relazione che lega parametro marginale e condizionato e che permette la scomposizione in effetti diretti ed effetti indiretti. Nell'ultima parte si è cercato di capire quando l'assunzione, effettuata molto spesso in letteratura, di un logit lineare nel modello marginale e nel modello ridotto possa ritenersi realistica.

Appendice A

Strumenti di algebra e probabilità

A.1 Il prodotto di Hadamard

Siano A e B due matrici di uguale dimensione $m \times n$, il *prodotto di Hadamard* tra A e B, indicato con $A \odot B$, torna una una matrice C di uguale dimensione tale che

$$C_{ij} = (A \odot B)_{ij} = A_{ij} \cdot B_{ij}, \quad 1 \leq i \leq m, 1 \leq j \leq n,$$

in generale si ha,

$$A \odot B = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \odot \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{bmatrix}$$

Per matrici di diverse dimensioni il *prodotto di Hadamard* non è definito.

A.2 La correlazione parziale

La correlazione parziale misura il grado di associazione tra due variabili casuali, eliminando l'effetto di un insieme di variabili definite di controllo. Sia quindi $U = (U_1, \dots, U_n)$ un vettore aleatorio di variabili casuali con matrice di varianza pari

a Σ. La correlazione parziale tra l'elemento i e l'elemento j di U è data da

$$\rho_{(i,j)\cdot\{U/(i,j)\}} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}},$$

dove σ^{ij} indica l'elemento in posizione (i, j) della matrice Σ^{-1}.

Nel caso di $n = 3$ variabili aleatorie, la formula per la correlazione parziale tra U_1 ed U_2 si riduce a

$$\rho_{(1,2)\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}.$$

La correlazione parziale coincide con la correlazione condizionata se il vettore aleatorio si distribuisce come una normale multivariata, come dimostrato in Baba et al. (2004).

A.3 La distribuzione logistica

La distribuzione logistica è una distribuzione di probabilità continua, con densità

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2},$$

per $x \in \mathbb{R}$, e viene indicata tramite $X \sim Lo(0, 1)$. Siano $\mu \in \mathbb{R}$ ed $s \in \mathbb{R}^+$ rispettivamente un parametro di locazione ed un parametro di scala, allora la variabile $Y = \mu + sX$ ha densità

$$f(y; \mu, s) = \frac{\exp(-\frac{x}{s})}{(1+\exp(-\frac{x}{s}))^2}$$

e viene indicata con $Y \sim Lo(\mu, s)$.

Sia $Y \sim Lo(\mu, s)$, allora la sua funzione generatrice dei momenti risulta

$$M_Y(t) = \exp(\mu t)B(1-st, 1+st) \quad t \in \left(-\frac{1}{s}, \frac{1}{s}\right),$$

dove $B(\cdot, \cdot)$ indica la funzione Beta. Ne deriva che

$$\mathbb{E}[Y] = \mu,$$
$$\mathbb{V}ar[Y] = s^2 \frac{\pi^2}{3}.$$

Appendice B

Studio di simulazione

B.0.1 Logit condizionato

In Tabella (B.1) per ogni valore di α e di n viene riportata:

- **Prima riga:** stima media dei $\hat{\beta}$ ottenuti tramite il metodo della massima verosimiglianza per il modello di regressione logistica utilizzando come covariate X e W;
- **Seconda riga:** frazione di volte in cui si accetta un test di ipotesi alla Wald di livello 5% su ogni singolo parametro: $\beta_0 = 0$, $\beta_x = \alpha_x$ e $\beta_w = \alpha_w$. Inoltre viene riportata la frazione di volte in cui si accetta l'ipotesi congiunta $H_0 : \beta_0 = 0, \beta_x = \alpha_x, \beta_w = \alpha_w$.

B.0.2 Logit marginale

In Tabella (B.2) per ogni α e per ogni n viene riportata:

- **Prima riga:** stima media dei $\hat{\beta}$ ottenuti tramite il metodo della massima verosimiglianza per il modello di regressione logistica utilizzando la sola X come covariata;
- **Seconda riga:** frazione di volte in cui si accetta un test di ipotesi alla Wald di livello 5% su ogni singolo parametro: $\beta_0 = 0$ e $\beta_x = \alpha_{x(x)}$. Inoltre viene riportata la frazione di volte in cui si accetta l'ipotesi congiunta $H_0 : \beta_0 = 0, \beta_x = \alpha_{x(w)}$.

α	$n = 250$				$n = 500$				$n = 1000$				$n = 3000$			
	β_0	β_x	β_w	H_0	β_0	β_x	β_w	H_0	β_0	β_x	β_w	H_0	β_0	β_x	β_w	H_0
$\alpha = (1, 0)$	-0.003	1.016	0.004	-	0.001	1.009	0.001	-	0.000	1.005	0.001	-	-0.000	1.002	-0.001	-
	0.950	0.955	0.947	0.955	0.948	0.953	0.953	0.949	0.947	0.952	0.947	0.947	0.951	0.951	0.948	0.951
$\alpha = (1, \frac{1}{4})$	0.001	1.025	0.256	-	0.002	1.007	0.252	-	0.001	1.003	0.252	-	0.000	1.002	0.250	-
	0.948	0.949	0.951	0.956	0.954	0.948	0.954	0.956	0.948	0.955	0.955	0.951	0.949	0.948	0.951	0.950
$\alpha = (1, \frac{1}{2})$	-0.003	1.019	0.513	-	-0.000	1.011	0.505	-	-0.001	1.003	0.504	-	-0.001	1.001	0.502	-
	0.953	0.953	0.952	0.955	0.950	0.953	0.954	0.955	0.953	0.952	0.949	0.955	0.948	0.948	0.955	0.950
$\alpha = (1, 1)$	-0.002	1.027	1.026	-	-0.000	1.008	1.012	-	0.002	1.007	1.003	-	0.001	1.002	1.003	-
	0.953	0.954	0.952	0.959	0.947	0.948	0.948	0.948	0.952	0.955	0.948	0.950	0.952	0.950	0.953	0.949
$\alpha = (1, 2)$	-0.003	1.027	2.060	-	-0.000	1.012	2.030	-	-0.000	1.005	2.016	-	0.000	1.001	2.005	-
	0.953	0.951	0.951	0.956	0.948	0.947	0.956	0.954	0.948	0.951	0.946	0.946	0.953	0.953	0.951	0.951
$\alpha = (1, 4)$	-0.004	1.043	4.185	-	-0.000	1.018	4.086	-	-0.001	1.011	4.043	-	-0.001	1.006	4.016	-
	0.954	0.945	0.958	0.957	0.952	0.956	0.957	0.957	0.948	0.952	0.947	0.951	0.951	0.948	0.944	0.947
$\alpha = (1, -\frac{1}{4})$	0.002	1.023	-0.256	-	0.001	1.008	-0.252	-	0.000	1.005	-0.252	-	-0.000	1.001	-0.250	-
	0.946	0.945	0.953	0.952	0.951	0.949	0.956	0.955	0.949	0.948	0.952	0.950	0.952	0.949	0.946	0.951
$\alpha = (1, -4)$	0.004	1.034	-4.132	-	0.001	1.017	-4.074	-	0.001	1.006	-4.025	-	0.002	1.005	-4.014	-
	0.953	0.956	0.959	0.956	0.945	0.952	0.950	0.954	0.948	0.951	0.955	0.951	0.954	0.955	0.948	0.953
$\alpha = (-3, -\frac{1}{4})$	0.001	-3.105	-0.255	-	0.002	-3.052	-0.250	-	0.000	-3.023	-0.253	-	0.002	-3.007	-0.250	-
	0.953	0.957	0.949	0.953	0.952	0.948	0.953	0.950	0.951	0.949	0.949	0.950	0.947	0.949	0.945	0.951
$\alpha = (-3, 4)$	-0.002	-3.102	4.135	-	0.000	-3.051	4.066	-	0.001	-3.032	4.039	-	-0.001	-3.007	4.007	-
	0.951	0.952	0.956	0.955	0.951	0.951	0.955	0.954	0.952	0.958	0.954	0.955	0.949	0.951	0.952	0.946

Tabella B.1: Regressione logistica di Y su X e W, con $\bar{\omega}_{xw} = 0.5$ e numerosità Monte-Carlo pari a 5000.

α	$n = 250$			$n = 500$			$n = 1000$			$n = 3000$		
	β_0	β_x	H_0	β_0	β_x	H_0	β_0	β_x	H_0	β_0	β_x	H_0
$\alpha = (1, 0)$	-0.003	1.013	-	0.001	1.007	-	0.000	1.004	-	-0.000	1.001	-
	0.950	0.958	0.954	0.951	0.951	0.946	0.948	0.951	0.944	0.952	0.954	0.948
$\alpha = (1, \frac{1}{4})$	0.001	1.136	-	0.002	1.120	-	0.001	1.117	-	0.000	1.116	-
	0.950	0.949	0.950	0.955	0.949	0.953	0.949	0.954	0.950	0.950	0.948	0.951
$\alpha = (1, \frac{1}{2})$	-0.004	1.222	-	-0.000	1.214	-	-0.001	1.207	-	-0.000	1.205	-
	0.955	0.954	0.952	0.951	0.949	0.952	0.956	0.950	0.954	0.952	0.942	0.946
$\alpha = (1, 1)$	-0.001	1.342	-	-0.001	1.325	-	0.002	1.322	-	0.001	1.319	-
	0.953	0.948	0.946	0.948	0.933	0.939	0.950	0.933	0.940	0.952	0.903	0.911
$\alpha = (1, 2)$	-0.002	1.392	-	-0.001	1.375	-	-0.001	1.371	-	-0.000	1.368	-
	0.951	0.929	0.928	0.952	0.902	0.915	0.951	0.863	0.890	0.949	0.704	0.759
$\alpha = (1, 4)$	-0.002	1.310	-	-0.001	1.298	-	0.001	1.294	-	-0.001	1.291	-
	0.955	0.913	0.921	0.948	0.880	0.896	0.949	0.799	0.834	0.954	0.539	0.632
$\alpha = (1, -\frac{1}{4})$	0.002	0.882	-	0.001	0.871	-	0.000	0.869	-	-0.000	0.867	-
	0.947	0.946	0.948	0.952	0.953	0.955	0.950	0.951	0.950	0.953	0.952	0.951
$\alpha = (1, -4)$	0.002	-0.419	-	-0.000	-0.420	-	0.001	-0.416	-	0.001	-0.414	-
	0.950	0.936	0.939	0.949	0.916	0.967	0.945	0.888	0.903	0.949	0.733	0.794
$\alpha = (-3, -\frac{1}{4})$	0.000	-3.184	-	0.002	-3.142	-	0.000	-3.120	-	0.002	-3.105	-
	0.955	0.956	0.955	0.952	0.948	0.951	0.952	0.951	0.951	0.948	0.950	0.950
$\alpha = (-3, 4)$	-0.003	-0.420	-	0.003	-0.418	-	0.000	-0.417	-	-0.001	-0.415	-
	0.950	0.937	0.941	0.954	0.916	0.928	0.953	0.887	0.911	0.946	0.751	0.805

Tabella B.2: Regressione logistica di Y su X, con $\bar{\omega}_{xw} = 0.5$ e numerosità Monte-Carlo pari a 5000.

Bibliografia

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, *59*(1), 19–35.

Arnold, B. C. & Beaver, R. J. (2000). Hidden truncation models. *Sankhyā: The Indian Journal of Statistics, Series A*, 23–35.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A. & Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, *58*(3), 471–488.

Azzalini, A. (1985). A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Azzalini, A. & Capitanio, A. (1999). Statistical Applications of the Multivariate Skew Normal Distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *61*(3), 579–602.

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*(4), 715–726.

Azzalini, A. & Capitanio, A. (2013). *The Skew-Normal and Related Families*. Cambridge University Press.

Baba, K., Shibata, R. & Sibuya, M. (2004). Partial correlation and conditional correlation as measure of conditional independence. *Australian & New Zealand Journal of Statistics*, *46*(4), 657–664.

Baron, R. M. & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Capitanio, A., Azzalini, A. & Stanghellini, E. (2003). Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, *30*, 129–144.

Caravenna, F. & Dai Pra, P. (2013). *Probabilità: Un'introduzione attraverso modelli e applicazioni*. Springer Milan.

Cochran, W. G. (1938). The Omission or Addition of an Independent Variate in Multiple Linear Regression. *Supplement to the Journal of the Royal Statistical Society*, *5*, 171–176.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

de Helguero, F. (1908). Sulla rappresentazione analitica delle curve abnormali. *Atti del IV Congresso Internazionale dei Matematici*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.

Greenland, S., Pearl, J. & Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical science*, *14*(1), 29–46.

Iacobucci, D. (2008). *Mediation analysis*. Sage.

Karlson, K. B., Holm, A. & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological methodology*, *42*(1), 286–313.

Lin, D. Y., Psaty, B. M. & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 948–963.

MacKinnon, D., Warsi, G. & Dwyer, J. (1995). A simulation study of mediated effect measures. *Multivariate behavioral research*, *30*, 41.

MacKinnon, D. P. (2012). *Introduction to statistical mediation analysis*. Routledge.

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

Neuhaus, J. M. & Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, *80*(4), 807–815.

Salvan, A., Sartori, N. & Pace, L. Modelli lineari generalizzati. Springer, 2020.

Serfling, R. (2004). Multivariate Symmetry and Asymmetry.

Stanghellini, E. & Doretti, M. (2019). On marginal and conditional parameters in logistic regression models. *Biometrika*, *106*(3), 732–739.

VanderWeele, T. J. (2015). *Explanation in causal inference : methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J. (2016). Mediation analysis: a practitioner's guide. *Annual review of public health*, *37*, 17–32.

VanderWeele, T. J. & Vansteelandt, S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*, *172*(12), 1339–1348.

Winship, C. & Mare, R. D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology*, *89*(1), 54–110.

Xie, X., Ma, Z. & Geng, Z. (2008). Some association measures and their collapsibility. *Statistica Sinica*, 1165–1183.

Zacks, S. (1981). *Parametric statistical inference: basic theory and modern approaches* (Vol. 4). Elsevier.

Ringraziamenti

In questa ultima parte, desidero ringraziare tutte le persone mi hanno accompagnato in questo percorso di crescita professionale e personale.

In primo luogo, vorrei ringraziare il Professor Scarpa e la Professoressa Stanghellini per avermi supportato ed aiutato in questo lavoro di tesi, oltre che per la loro disponibilità ed i loro preziosi consigli. Ringrazio inoltre il Prof. Aliverti per le chiacchierate e per i suggerimenti su corsi e lezioni.

Un ringraziamento doveroso va alla mia famiglia, ed in particolare ai miei genitori, che mi hanno accompagnato e sostenuto durante questo percorso, incoraggiandomi a dare il meglio di me stesso e a non mollare nei momenti di difficoltà.

Ringrazio inoltre gli amici di sempre e tutte le persone con cui ho condiviso questo fantastico viaggio durato 5 anni. Un ringraziamento speciale va a Marco, Alessia, Pietro e Riccardo che mi hanno sempre appoggiato e con cui ho trascorso gran parte di questo percorso.

Alle volte uno si crede incompleto
ed è soltanto giovane.
(Italo Calvino)