Università degli studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale
in Scienze Statistiche

Analisi delle mortalità per causa attraverso i modelli GLM Zero Inflated

Relatore: Prof. Stefano Mazzuco
Dipartimento di Scienze Statistiche

Laureanda: Mara Giacon
Matricola n. 1098288

Anno Accademico 2016/2017

Indice

Introduzione

Questo elaborato si occuperà di studiare la mortalità per causa [Missov e Lenart 2016], i dati si riferiscono alla Francia e riguardano gli anni dal 2000 al 2013. L'interesse per questo tipo di analisi è dettato da motivazioni socio-demografiche, infatti ogni Paese è interessato a carpire eventuali fattori di rischio che possano mettere a repentaglio l'incolumità dei propri cittadini e la possibilità di attuare politiche adeguate per prevenirli e combatterli. Interessante è anche studiare le dinamiche legate allo scorrere del tempo e le eventuali modifiche che intercorrono nello stile di vita dei viventi.

Nel dettaglio, si vedrà nel primo capitolo una rassegna sulle principali motivazioni per cui si studia la mortalità per causa e la nascita di questa tematica come oggetto di analisi statistica.

Nel secondo capitolo verranno presentati i dati, le informazioni di cui si dispone e le principali azioni che vengono compiute per poterli usare. Sarà presente anche una descrizione degli organi che hanno reperito e reso disponibile il *dataset* e il lavoro che hanno intrapreso per ottenere queste raccolte, così preziose e accurate. Sempre in questo capitolo, è inserita una descrizione dei dati utilizzati e tutta una serie di altre informazioni necessarie in caso possa risultare interessante procedere con altre analisi relative ad altri paesi o ad altri livelli di dettaglio nel trattamento delle cause. Una sezione di questo capitolo sarà dedicata ad un'analisi iniziale esplorativa dei dati disponibili per meglio comprendere la situazione generale.

Il terzo capitolo conterrà una spiegazione teorica dei modelli che verranno utilizzati per analizzare i dati, in particolare sono descritti i modelli lineari generalizzati, con risposta di Poisson e Binomiale Negativa, nella prima parte, e i modelli ad inflazione di zeri per le stesse distribuzioni, nella seconda. Si è fatta particolare attenzione

alla descrizione sia dell'implementazione dei modelli che a quella relativa alle analisi grafiche necessarie per poter giudicare l'accuratezza di ciascuno di essi.

Le varie implementazioni di questi modelli ai dati sono descritte nel capitolo 4, dove si potranno vedere sia le stime che si sono ottenute, che le analisi diagnostiche circa le criticità e le potenzialità di ciascun modello. Per ciascuno dei modelli utilizzati si presenterà una sola combinazione delle variabili disponibili, quella ritenuta più adeguata. In questo capitolo non sono state inserite tutte le informazioni ricavate da tali implementazioni per una questione di pesantezza del lavoro esposto, alcune di queste sono presenti in appendice alla fine dell'elaborato. Questo capitolo si conclude con un confronto tra i vari modelli atto a comprendere quale tra quelli utilizzati sia quello che meglio si adatta ai dati disponibili e un veloce confronto tra le stime dei diversi modelli scelti per uomini e donne, con relativa interpretazione.

Il quinto e ultimo capitolo tratterà un'estensione di uno dei modelli presentati nei capitoli 3 e 4, per cercare di meglio assecondare le dinamiche intrinseche ai dati di cui si dispone. Si tratta di un'estensione del modello ad inflazione di zeri con risposta Binomiale Negativa per dati gerarchici. In questo capitolo sono presenti sia delle delucidazioni teoriche e descrittive, sia l'implementazione del modello analizzato ai dati con le relative analisi diagnostiche.

Capitolo 1

Perchè si studia la mortalità per causa

Dalla notte dei tempi, alla morte è stata data notevole importanza, questo evento, che determina la fine della vita di ogni individuo, è infatti circondato da mistero e venerazione. Sin dall'antichità si sono susseguite leggende e riti che accompagnavano questo momento; tant'è che molte religioni si basano su questo accadimento per professare le più svariate teorie circa le ragioni e il postumo di esso.

Anche in letteratura uno dei temi più dibattuti e più trattati è quello relativo alla morte: ogni autore, filosofo, artista, letterato, carica questo evento di significati e interpretazioni personali differenti. Solo per citarne alcuni tra i più famosi e importanti che hanno trattato questo argomento si possono menzionare Alessandro Manzoni, Giacomo Leopardi, Ugo Foscolo, Eugenio Montale, Victor Hugo, Arthur Schopenhauer, Oscar Wilde, ma sono veramente moltissimi coloro che si potrebbero nominare.

Anche nei tempi più moderni questo evento viene caricato di importanza e trattato negli ambiti più disparati. I dati che riguardano questo aspetto della vita sono molteplici e in particolare la causa che ha provocato la morte dell'individuo è uno dei più interessanti che si possano studiare. Questo aspetto, non è però solo stimolante da un punto di vista socio-demografico ma anche statistico, in particolare attorno a questo evento si possono trarre molteplici informazioni utili per il benessere degli esseri umani sia a livello nazionale che internazionale. Concentrandosi sullo studio statistico di questo fenomeno, si vedranno ora alcuni aspetti che possono risultare utili per comprendere come, la materia analizzata, determini tale interesse.

Lo studio di questo evento può diventare anche una necessità sotto determinati aspetti: i Paesi sono infatti interessati a controllare le cause di morte e la loro evoluzione nel tempo al fine di tenere sotto controllo eventuali cambiamenti che intercorrono e che colpiscono i cittadini. Tali rilevazioni e lo studio di questi dati può infatti permettere di prevenire l'insorgere di epidemie o fermare comportamenti dannosi. Queste abitudini fanno parte della vita di chiunque, e grazie ad un'accurata informazione in materia e a dei mutamenti, talvolta radicali, nelle usanze di ciascuno, si potrebbe riuscire a ridurre l'incidenza di alcune cause; esempi di questi atteggiamenti possono essere l'uso di alcool, di droghe e il fumo, un regime alimentare scorretto, magari legato anche alla sedentarietà e alla mancanza di una cultura legata all'attività fisica o la mancata prevenzione per alcuni tipi di malattie che si potrebbe attuare grazie a periodiche visite mediche, ma ci sono anche comportamenti scorretti da parte di aziende, associazioni o enti preposti, come il trattamento inadeguato di sostanze tossiche o lo smaltimento sbagliato dei rifiuti, consapevole o meno che sia. Torna utile sfruttare l'analisi di questi dati per l'educazione degli abitanti e per attuare campagne di sensibilizzazione verso abitudini errate che possano mettere in pericolo la vita degli stessi o della popolazione in generale. A questo scopo possono essere condotti anche studi specifici che considerino non solo l'evoluzione nel tempo dei tassi di morte per causa, ma anche nello spazio, in modo da avere una visione più completa delle dinamiche di questo fenomeno.

Gli utilizzatori dei dati sulle cause di morte possono essere le persone più disparate [Simpson 2017b]:

- istituzioni nazionali e internazionali, come i Governi dei Peasi, gli Enti locali, la Commissione Europea, l'Organizzazione mondiale della sanità o altre associazioni; lo scopo di questi Istituti è generalmente legato alla prevenzione e alla salvaguardia della salute dei cittadini, al miglioramento dell'andamento della vita degli abitanti. Grazie a questi studi possono infatti prendere decisioni consapevoli che riguardano le politiche pubbliche, in materia di salute, e sanitarie, sia a livello di prevenzione di comportamenti scorretti che di analisi degli effetti di determinate decisioni prese in passato, sia alla possibilità di stanziare fondi al fine di ridurre, ove possibile, l'incidenza di alcune cause;

- ricercatori di statistica, in particolare gli Istituti nazionali di statistica dei vari Paesi, come ad esempio per l'Italia l'Istat o a livello di Comunità Europea l'Eurostat: per rendere possibili confronti a livello territoriale e temporale, e far sì che le informazioni fornite siano di qualità superiore, così da ottenere analisi più accurate possibile;
- piccoli enti pubblici, aziende private o singoli ricercatori interessati all'argomento: le motivazioni che spingono costoro a utilizzare questi dati possono essere le più diverse, da semplice interesse personale motivato dalla voglia di conoscere in modo più approfondito possibile il fenomeno, per il singolo ricercatore, a potenziale mercato di investimento ed espansione, per le aziende private, alla ricerca di modi per fare prevenzione per zone circoscritte per i piccoli enti pubblici e attuare politiche socio-sanitarie *ad hoc*;
- come supporto in indagini e processi sull'insorgenza di patologie e sulla proliferazione di epidemie da parte di tribunali e procure, per la ricerca di eventuali responsabilità in merito a casi di inquinamento ambientale o negligenza nelle corrette procedure di trattamento di sostanze tossiche o dannose, che possano minare la salute degli abitanti;
- i media, dai giornali alle trasmissioni televisive che quotidianamente trattano questo tipo di argomento per diversi aspetti, sia a livello nazionale, che europeo o internazionale: ormai è stato ribadito più volte come questo argomento sia così interessante per molti cittadini, per l'appunto come già citato, essendo coperto da un velo di mistero coniugato allo stesso tempo a sensazioni di timore, che accompagna questo momento della vita. Quale, quindi, miglior specchietto per le allodole per catturare l'attenzione della maggior parte del pubblico? Le argomentazioni con cui viene trattato questo argomento dai media sono i più disparati, anche se non sempre attendibili: da inchieste condotte dai giornalisti stessi, a conclusioni a cui sono giunti traendo ispirazione da una o dall'altra pubblicazione.

Nel seguito dell'elaborato sarà presentata un'analisi dei dati relativi alle cause di mortalità, con particolare attenzione all'infuenza di diversi fattori, come la classe di età e il sesso degli individui.

Capitolo 2

Human Cause of Death e analisi esplorative

2.1 I dati

I dati contenuti ed analizzati in questo elaborato sono stati reperiti da *The Human Cause-of-Death Dataset*, all'indirizzo `http://www.causesofdeath.org/cgi-bin/main.php`. Questa raccolta di informazioni ha avuto origine da un progetto sorto dalla collaborazione tra il parigino *the French Institute for Demographic Studies (INED)* e il tedesco *Max Planck Institute for Demographic Research (MPIDR)*, sito a Rostock [*Human Cause-of-Death Database*].

Tali informazioni sono reperibili gratuitamente e ad accesso libero, è sufficiente registrarsi per potervi accedere; sono state raccolte per documentare l'andamento di alcune cause di morte e malattie, per poter quindi facilitare eventuali analisi a riguardo.

Sono disponibili i dati relativi ad un totale di 16 nazioni, europee e non, i quali contengono le numerosità di individui deceduti per una delle cause considerate, fissate a priori, per genere, ad una determinata età e in uno degli anni presi in esame. Questi ultimi risultano diversi da stato a stato: per alcuni, come la Lettonia, sono disponibili le informazioni di mezzo secolo, mentre per altri, come il Giappone o gli Stati Uniti, di poco più di una decina di anni. Il lavoro per la ricostruzione di

serie storiche in grado di considerare un più ampio arco temporale è tuttora in fase di sviluppo.

Per quanto concerne l'età degli individui, la stessa è raccolta in modo aggregato, in un numero di classi che può arrivare in alcuni casi anche a 26. Si tratta di intervalli prevalentemente quinquennali, eccezion fatta per la prima classe, la quale rappresenta coloro che sono nel primo anno di vita, ossia dalla nascita ad un anno di età non ancora compiuto, e per alcune altre classi che risultano aperte a destra[1]. Queste ultime sono rispettivamente 85, 90, 95, e 100 anni; esse però non sono presenti in tutti gli stati e per tutti gli anni considerati, infatti in alcuni casi non si tratta di un'informazione disponibile e le classi di età considerate si riducono a 24 (nel caso l'ultima classe sia rappresentata dai 95 anni e oltre), 22 (nel caso di 90 e oltre) oppure 20 (con 85 e oltre).

Per le cause di morte sono considerate tre diverse specificazioni: la *Short list* che contiene 16 cause, la *Intermediate list* che ne contempla 104 e la *Full list* che rappresenta la classificazione maggiormente dettagliata, pur tuttavia diversa da stato a stato: in questa suddivisione per categorie le cause considerate sono oltre 4700. Punto fermo nell'elaborazione delle serie storiche presenti è stata senza dubbio la comparazione delle informazioni raccolte in anni diversi e per i molteplici paesi, nonostante nel tempo siano intervenute diverse modifiche nelle varie classificazioni ed esistano procedure nazionali differenti: questo al fine di garantire omogeneità, per quanto possibile, nei risultati che si possono ottenere sottoponendo i dati ad un'analisi approfondita. A tal proposito, sul piano delle prime due rappresentazioni, sono stati resi disponibili *dataset* aggiuntivi che considerano un'ulteriore causa per la *Short list* e due per la *Intermediate list*, portandole ad un totale di 17 per la prima e 106 per la seconda.

Il lavoro espletato sulla *Full list* invece è rappresentato dalla costruzione di coefficienti di transizione per poter uniformare i dati raccolti con diverse classificazioni,

[1]Per "classi aperte a destra" si intendono intervalli nei quali è indicato il solo estremo inferiore mentre quello superiore è lasciato variare in base al massimo presente, come ad esempio definendo la classe [87 e oltre) si intendono compresi in questa tutti coloro che vanno dal 87-esimo anno di età compiuto, fino al massimo presente nel campione in quel momento, questo può variare nel tempo in base agli individui in essere

grazie ad un metodo implementato da INED negli anni '80[2] che in questo contesto è stato comunque personalizzato per i singoli Paesi e le particolari necessità intrinseche a ciascuno di essi. In tutte e tre le classificazioni le cause sono esaustive ed esclusive.

Per ciascuna della caratteristiche considerate oltre alle modalità già presentate è disponibile il totale di classe.

Sono reperibili non solo i conteggi degli individui morti per una determinata causa, ma anche i dati relativi alla popolazione totale di riferimento, vale a dire la popolazione a rischio, esposta alla possibilità di incorrere in quella causa di morte, oltre ai tassi grezzi e standardizzati e alla numerosità dei nati; tutte queste informazioni sono sempre riportate suddivise per genere e fascia di età.

2.2 Il *dataset*

Per la stesura di questo elaborato è stato scelto il *dataset* relativo alla Francia, dove sono disponibili i dati relativi agli anni dal 2000 al 2013. La scelta di questo Paese è dovuta alla buona qualità dei dati e all'adeguatezza delle rilevazioni. La lista delle cause usata è quella breve che comprende [*Human Cause-of-Death Database*]:

- causa 1: malattie infettive;
- causa 2: neoplasma;
- causa 3: malattie del sangue e degli organismi sanguigni;
- causa 4: malattie endocrine, nutrizionali e metaboliche;
- causa 5: disturbi mentali e comportamentali;
- causa 6: malattie del sistema nervoso e degli organi di senso;
- causa 7: malattie cardiache;
- causa 8: malattie cerebrovascolari;
- causa 9: altri disturbi e disturbi non specificati del sistema circolatorio;
- causa 10: malattie respiratorie acute;

[2]Il metodo in esame comprende tre fasi tra le quali un test statistico

- causa 11: altre malattie respiratorie;
- causa 12: malattie del sistema digerente;
- causa 13: malattie della pelle e del tessuto sottocutaneo, del sistema muscolo-scheletrico e del tessuto connettivo;
- causa 14: malattie del sistema genitourinario e complicanze della gravidanza, del parto e del puerperio;
- causa 15: alcune condizioni originarie del periodo perinatale e anomalie o malformazioni congenite;
- causa 16: cause esterne;
- causa 17: altre malattie o causa ignota.

Sul piano delle analisi, esse sono state condotte separatamente per uomini e donne: questo perchè è risaputo che vi siano dinamiche differenti che influenzano in modo diverso gli individui in base al genere. Rispetto alle età, malgrado sia stato provato ad aggregare in diversi modi le varie classi, non sono stati raggiunti sensibili miglioramenti sullo studio sviluppato; si sono quindi quasi totalmente mantenute le classi reperite direttamente dal *database*, in particolare si sono usate: [0-1), [1-5), [5-10)], [10-15), [15-20), [20-25), [25-30), [30-35), [35-40), [40-45), [45-50), [50-55), [55-60), [60-65), [65-70), [70-75), [75-80), [80-85), [85-90), [90-95), [95 e oltre).

Il *database* su cui sono state condotte le analisi trae vigore da tre *dataset*:

- *FRA_d_short_idr.csv*: contenente le numerosità degli individui francesi morti per una della prime 16 cause, divisi in 26 classi di età, per genere e per anno (dal 2000 al 2013) e altre informazioni riguardanti la descrizione dei dati come lo stato, la tipologia di classificazione per l'età e il tipo di dettaglio della lista delle cause;
- *FRA_d_idshort.csv*: contenente le numerosità degli individui francesi morti per la causa 17, divisi in 26 classi di età, per genere e per anno;
- *FRA_e.csv*: contenente le quantità della popolazione francese di riferimento, divisa in 24 classi di età, per genere e per anno.

Dopo numerose operazioni di sistemazione del *dataset* effettuate con il *software* *R*, disponibile gratuitamente all'indirizzo `http://www.r-project.org/`, si è ottenuto

un *dataset* contenente 9996 osservazioni e 7 variabili (successivamente diviso in due da 4998 individui ciascuno, uno per le donne e uno per gli uomini), ossia:

- *anno*: variabile numerica intera che indica l'anno della rilevazione, questa variabile è stata riscalata rispetto al suo minimo (ossia togliendo 2000 a tutti i valori) per meglio procedere con l'interpretazione dei parametri (va quindi da 0 a 13, anzichè da 2000 a 2013);
- *sesso*: variabile categoriale dicotomica che indica il sesso degli individui (uomo: M o donna: F);
- *cause*: variabile categoriale che indica la causa del decesso (da 1 a 17);
- *età*: variabile categoriale che indica la classe di età a cui appartengono gli individui (da 2 a 22);
- *morti*: variabile numerica intera che indica quanti individui con determinate caratteristiche sono deceduti (il cui minimo è pari a 0 mentre il massimo a 17390);
- *popolazione*: variabile numerica intera che indica il numero di individui con determinate caratteristice che sono presenti nella popolazione di riferimento, ossia a rischio di morire per una delle possibili cause (il cui minimo è pari a 14060 mentre il massimo è 2237000);
- *proporzione*: variabile continua che indica la proporzione di individui con determinate caratteristiche deceduti per una specifica causa, è stata calcolata come rapporto tra i morti per una causa con specifiche caratteriche e la popolazione di riferimento a rischio (il cui minimo è 0 mentre il massimo è 0.119).

2.3 Analisi esplorative

In questa parte verranno analizzate con maggior dettaglio, le osservazioni disponibili in base alle variabili presenti nel *dataset*, per meglio comprendere la composizione dei dati e la situazione che si presenta. Nella sezione precedente sono già state illustrate le caratteristiche principali delle singole variabili, come codifica, massimo, minimo, mentre ora si vedranno le combinazioni di queste applicate ai dati.

Tutte le analisi saranno distinte tra gli individui di sesso maschile e quelli femminili; infatti, come già poc'anzi anticipato, questa caratteristica risulta piuttosto distintiva. A conferma di tale assunzione si può vedere il grafico 2.1[3] dove appare evidente come questa suddivisione sia necessaria: si può infatti notare come le dinamiche che intercorrono nei due gruppi e le peculiarità intrinseche a ciascuno siano differenti.

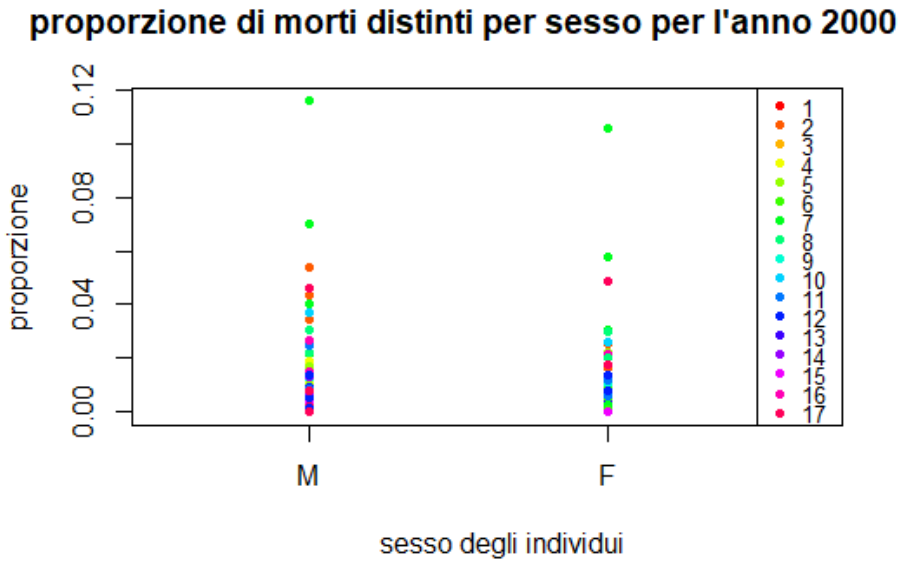

Figura 2.1: grafico della variabile proporzione divisa per sesso, relativo all'anno 2000: i diversi colori rappresentano le differenti cause di morte

Si prenderà ora in esame la variabile relativa alla classe di età di appartenenza degli individui, così come è stata descritta nella sezione precedente. I due grafici sottostanti 2.2 riportano le proporzioni di persone morte suddivisi per la classe di età di appartenenza, uno per gli uomini e uno per le donne.

Come si può notare le proporzioni per gli uomini tendono ad essere molto basse, eccetto che per la prima classe (relativa al primo anno di età), fino alla tredicesima classe di età (relativa all'intervallo 55-60 anni), e poi pian piano crescono in modo esponenziale, fino a raggiungere il massimo nell'ultima classe considerata (quella

[3]In questo caso non è stata rivolta particolare attenzione al tipo di causa (distinta per colore) del decesso o all'età degli individui, lo scopo di questo grafico è quello di vedere come siano differenti le composizioni delle proporzioni in base al sesso. Anche per quanto riguarda l'anno, è stato arbitrariamente scelto il primo a disposizione, ossia il 2000.

relativa ai 95 anni e oltre). Lo stesso grafico, ma riferito al sesso femminile, mostra un andamento anch'esso esponenziale: molto simile fino alla dodicesima classe (50-55 anni) a quello relativo agli uomini ma molto meno marcato per tutti gli intervalli successivi.

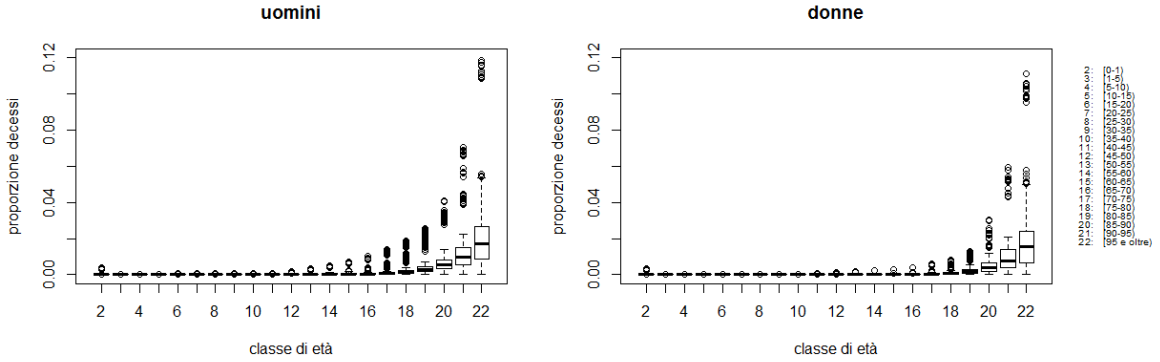

Figura 2.2: grafico della proporzione di decessi divisa per uomini e donne in funzione delle classi di età

Per meglio comprendere le dinamiche intrinseche al fenomeno di studio verranno ora proposti due grafici (2.3) simili ai precedenti ma che pongano maggiore attenzione sulle cause di morte.

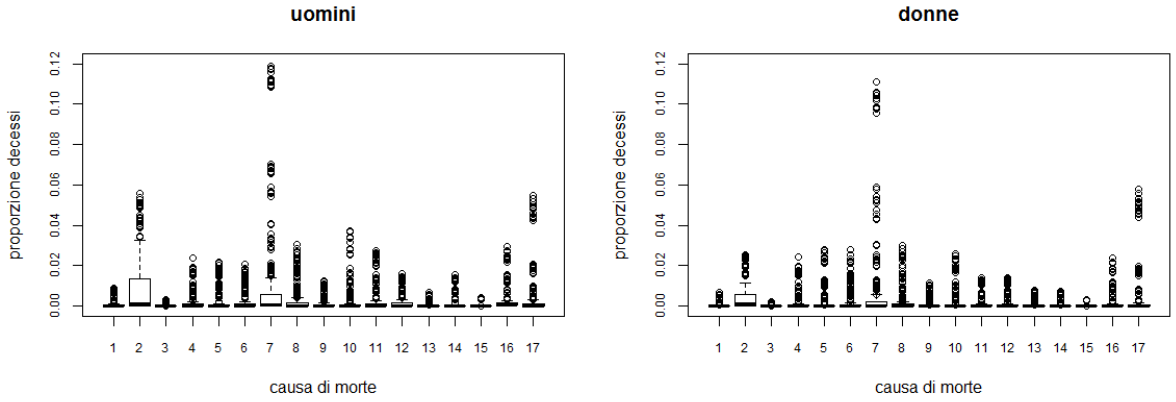

Figura 2.3: grafico della proporzione di decessi divisa per uomini e donne in funzione delle cause di morte

Analizzando le proporzioni di deceduti in funzione della causa di morte si può notare come per gli uomini la causa con valori maggiori è la 7, ossia le malattie cardiache, che arriva a circa un 12%, seguono con circa il 6% di valore massimo

le cause 2 e 17, vale a dire neoplasma e malattie non identificate. Valori piuttosto importanti sono raggiunti anche dalle cause 8, 10, 11 e 16, rispettivamente malattie cerebrovascolari, respiratorie e cause esterne. I valori più bassi si osservano per le cause 3 e 15, che interessano malattie del sangue, condizioni del periodo perinatale e anomalie congenite. Nel grafico relativo alle donne si può notare come, anche qui, la causa di morte con proporzioni maggiori sia la numero 7, ossia le malattie cardiache e, a seguire, la 17 che rappresenta una causa ignota. A differenza del grafico degli uomini non vi sono altre cause con proporzioni particolarmente accentuate ma vi è la presenza di 7 cause che raggiungono circa il 3% dei decessi ciascuna, rispettivamente: la 2 (neoplasma), la 4 (malattie endocrine, nutrizionali e metaboliche), la 5 (disturbi mentali e comportamentali), la 6 (malattie del sistema nervoso e degli organi di senso), la 8 (malattie cerebrovascolari), la 10 (malattie respiratorie acute) e la 16 (cause esterne). Anche per le donne le cause meno influenti sono la 3 (malattie del sangue e degli organismi sanguigni) e la 15 (alcune condizioni originarie del periodo perinatale e anomalie o malformazioni congenite).

I grafici fino ad ora descritti trasmettono un'idea complessiva delle dinamiche che influenzano il fenomeno studiato, il resto della sezione si concentra sul tentativo di capire con più precisione le peculiarità esistenti. Per rendere la comprensione più fluida, la variabile relativa alla classe di età di appartenenza è stata ulteriormente raggruppata in 5 macro classi (che verranno utilizzate solo per questi grafici):

- neonati: bambini entro il primo anno di età, non ancora compiuto;
- adolescenti: ragazzi che hanno un'età compresa tra 1 e 20 anni esclusi;
- giovani: coloro che hanno tra i 20 e i 40 anni non ancora compiuti;
- adulti: persone con un'età tra i 40 e i 70 anni esclusi;
- anziani: individui con più di 70 anni.

Nei grafici seguenti (dal 2.4 al 2.20) si vedrà l'andamento della proporzione di morti per la causa specifica, rispetto all'intervallo di 13 anni di dati analizzati e considerando le 5 nuove macro classi di età[4] sempre distinti per sesso.

[4] Ad eccezione di un solo caso la categoria con proporzioni maggiori è quella relativa alle persone anziane come si può notare da tutti i grafici, questa informazione non verrà ripetuta per ogni causa.

Dal grafico 2.4 della causa relativa alle malattie infettive, si può notare come la classe che ha una proporzione maggiore, quella di coloro che hanno più di 70 anni, abbia gli stessi andamenti oscillatori, con il passare degli anni, sia per gli uomini che per le donne: si potrebbe quindi supporre che esistano delle ragioni sottostanti che spieghino tale andamento. Le donne hanno proporzioni leggermente inferiori rispetto agli uomini per tutte le classi di età, eccetto che per i neonati, che, soprattutto dal 2001 al 2006, mostrano valori simili a quelli dei neonati maschi, se non addirittura superiori.

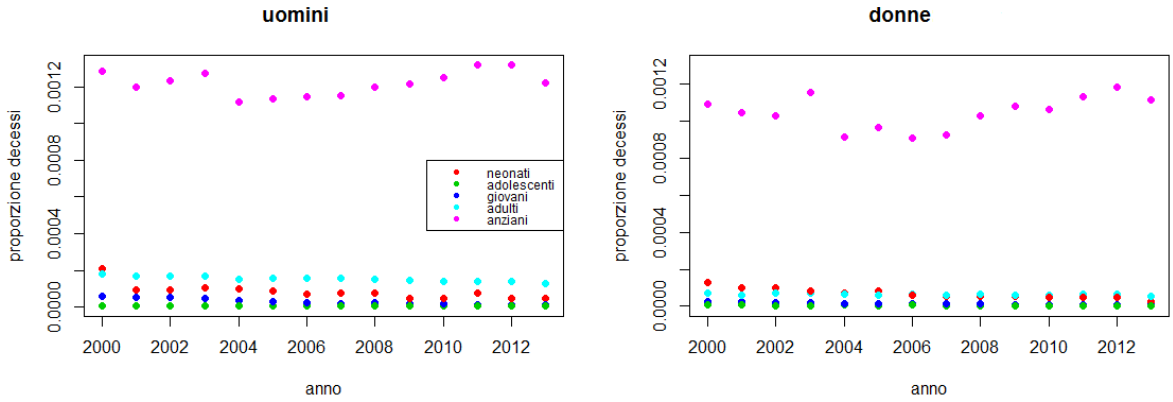

Figura 2.4: proporzione dei decessi per sesso, anno di rilevazione e classe di età, causa 1: malattie infettive

Nel grafico 2.5 viene analizzata la causa relativa alla morte da neoplasma, si vede che sia la classe di età con maggior incidenza, quella relativa alle persone anziane, sia la seconda, quella degli adulti, mostrano per gli uomini circa un valore doppio rispetto a quello delle donne. Però, mentre per i primi, con il passare degli anni, si può notare una tendenza a decrescere, per le donne resta costante. Già dal grafico 2.3 si poteva notare l'importanza di questa causa di morte per gli uomini.

I grafici 2.6, relativi alle malattie del sangue, mostrano come la situazione tra uomini e donne sia piuttosto simile rispetto ai valori delle proporzioni. Le donne mostrano però, con il passare del tempo, un'oscillazione più marcata per la categoria delle persone anziane, mentre negli uomini questa andatura meno lineare si può riscontrare nei neonati; la stessa classe per il genere femminile mostra valori più alti per i primi anni e poi valori quasi sempre molto vicini allo 0.

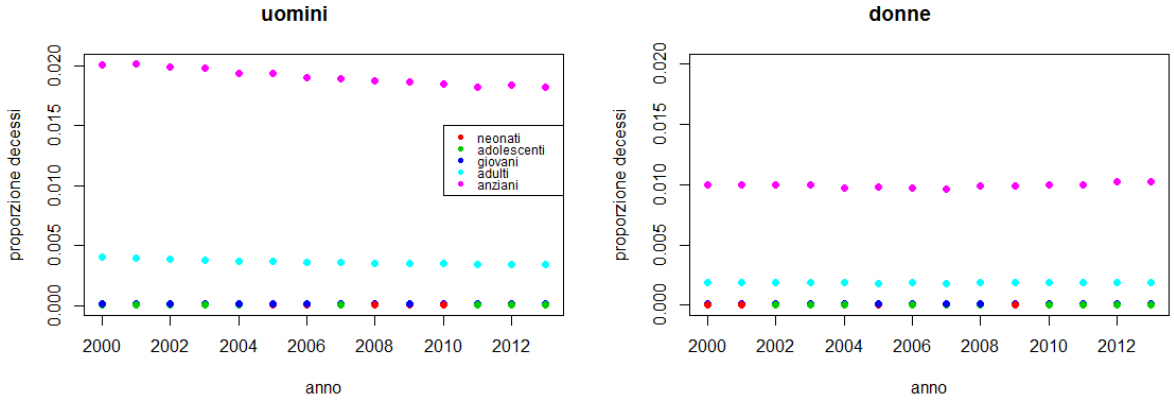

Figura 2.5: proporzione dei decessi per sesso, anno di rilevazione e classe di età, causa 2: neoplasma

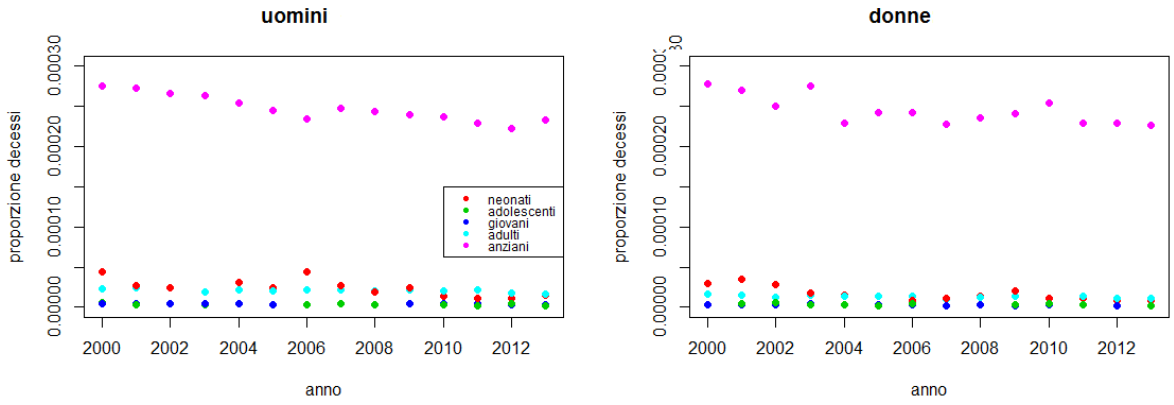

Figura 2.6: proporzione dei decessi per sesso, anno di rilevazione e classe di età, causa 3: malattie del sangue

Andamenti pressochè identici si possono osservare, tra uomini e donne, se si considera la causa di morte relativa a coloro che sono deceduti per malattie endocrine, nutrizionali o metaboliche, riportata nei grafici 2.7. L'unico scostamento tra i due sessi si riscontra nella categoria con proporzioni più elevate, quella degli anziani, le donne hanno valori leggermente più elevati ma in entrambi i sessi si può notare un picco nel 2003.

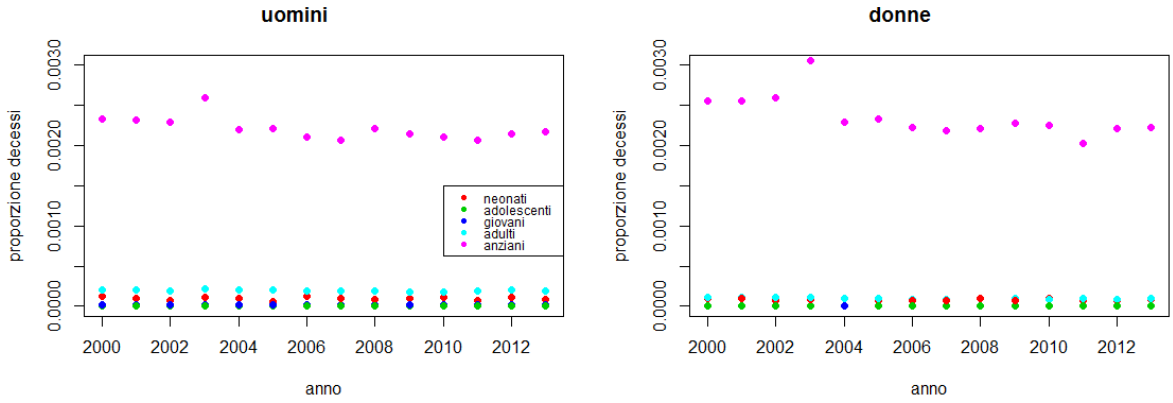

Figura 2.7: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 4: malattie endocrine, nutrizionali o metaboliche

Andamenti simili, ma a livelli differenti, anche se si considerano i grafici 2.8, relativi ai morti per disturbi mentali e comportamentali. La categoria delle persone anziane, quella maggiormente colpita, riscontra valori più alti per le donne. Si può notare come in questa classe di età, l'andamento sia inizialmente decrescente ma negli ultimi anni mostri un aumento. Se invece viene considerata la categoria degli adulti i valori più alti si riscontrano negli uomini, ciò potrebbe essere sintomo che questo genere di malattie colpisca in età più giovanile gli uomini rispetto alle donne.

Anche per il grafico 2.9, dove sono considerati i morti per malattie del sistema nervoso e degli organi di senso, si può rilevare un inesorabile aumento al decorrere degli anni, soprattutto nelle donne dove la proporzione di decessi per questo fattore è quasi raddoppiata passando dal 2000 al 2013; meno marcato è invece l'aumento per gli uomini anche se comunque importante.

Andamento opposto si ha invece nel grafico 2.10, dove sono considerati i morti per malattie cardiache. Si rileva un decremento sia per gli uomini che per le don-

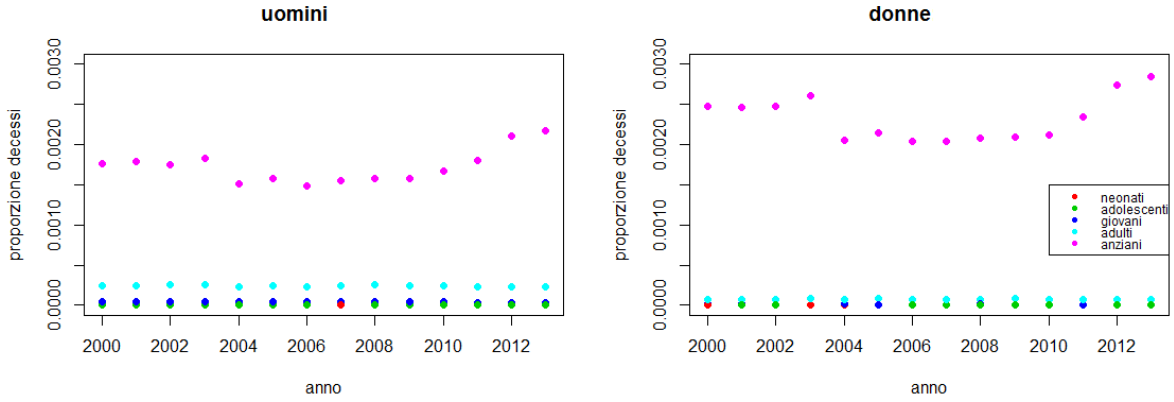

Figura 2.8: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 5: disturbi mentali e comportamentali

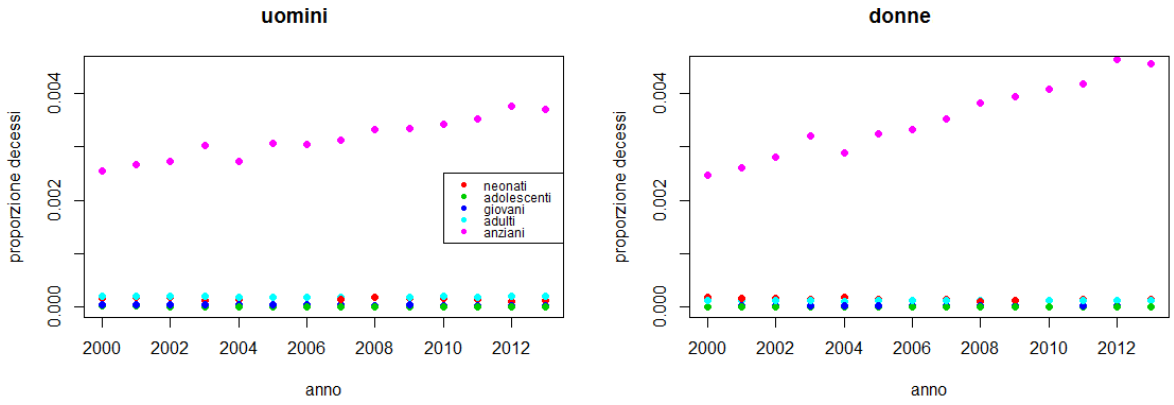

Figura 2.9: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 6: malattie del sistema nervoso e degli organi di senso

ne, forse dovuto ad una maggiore consapevolezza dell'importanza di una corretta alimentazione e di uno stile di vita attivo per prevenire questo tipo di patologie. Rispetto alla categoria degli adulti, gli uomini hanno proporzioni superiori rispetto alle donne, anche se di poco.

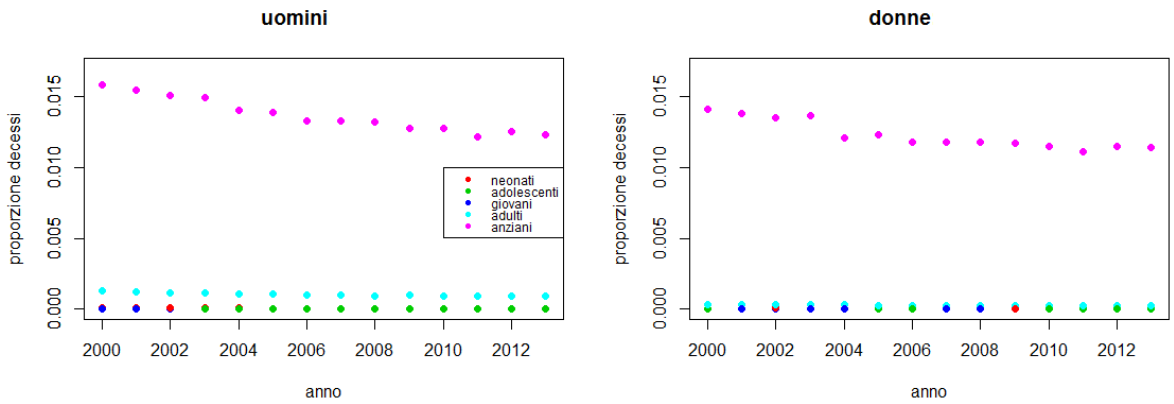

Figura 2.10: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 7: malattie cardiache

Anche per le malattie cerebrovascolari, si può notare dal grafico 2.11 un decremento nel corso del tempo della proporzione di morti per questa causa. Ancora una volta la categoria degli uomini adulti denota proporzioni leggermente più elevate rispetto alla medesima categoria femminile.

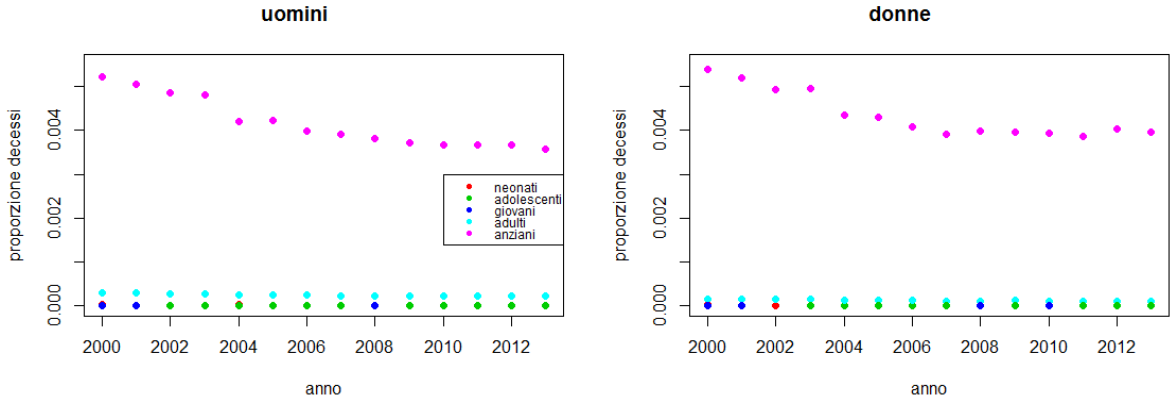

Figura 2.11: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 8: malattie cerebrovascolari

Il decremento per i morti per altri disturbi dell'apparato circolatorio, come si

può notare nel grafico 2.12, è piuttosto marcato, soprattutto per gli uomini, dove la proporzione quasi dimezza negli anni.

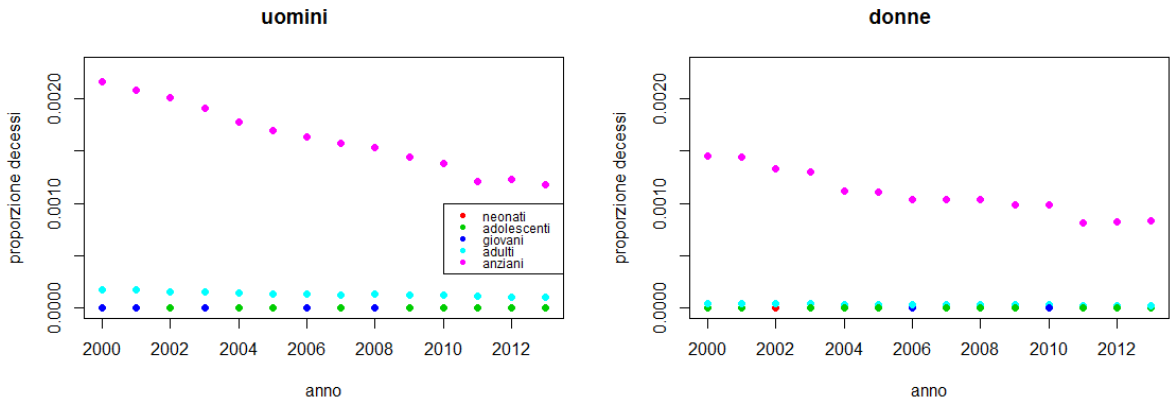

Figura 2.12: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 9: altri disturbi dell'apparato circolatorio

L'andamento oscillatorio presente nel grafico 2.13, che riporta i morti per malattie respiratorie acute, è molto simile per uomini e donne, nonostante i primi abbiano valori generalmente più elevati.

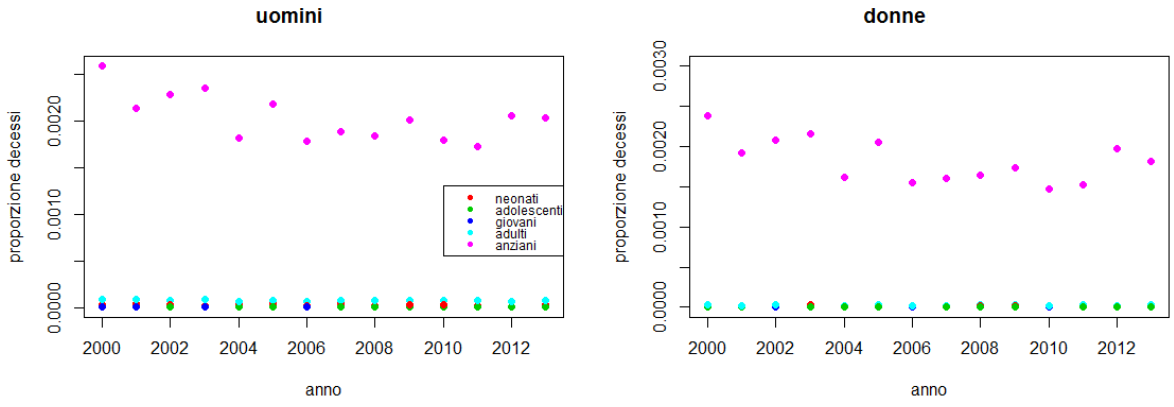

Figura 2.13: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 10: malattie respiratorie acute

Molto più apprezzabile è la differenza tra uomini e donne nel grafico 2.14, che riguarda i morti per altri tipi di malattie respiratorie (non acute) dove i primi manifestano valori circa doppi rispetto alle donne.

Molto simili sono invece le proporzioni di morti per le malattie del sistema digerente, visibili nel grafico 2.15, per la categoria delle persone anziane, mentre più

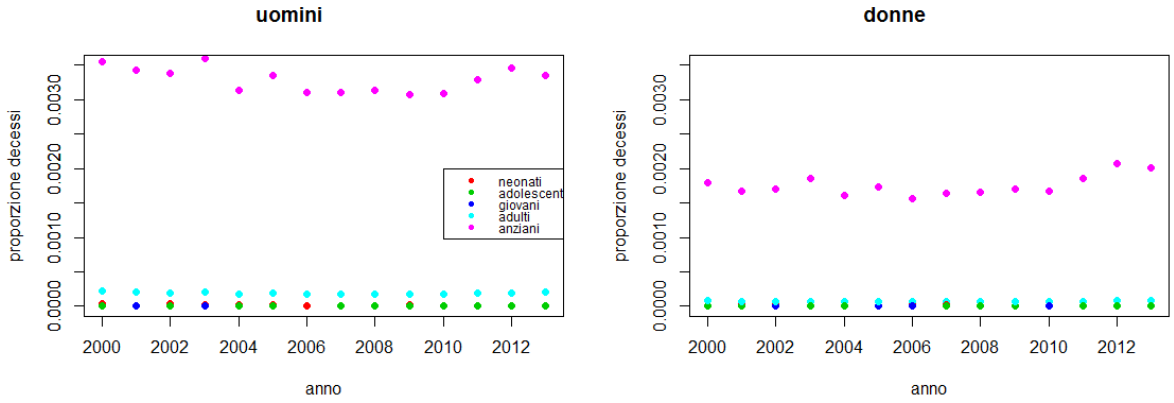

Figura 2.14: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 11: altre malattie respiratorie non acute

alti, circa doppi, sono i valori per gli adulti uomini rispetto alla stessa categoria di età ma riferita alle donne.

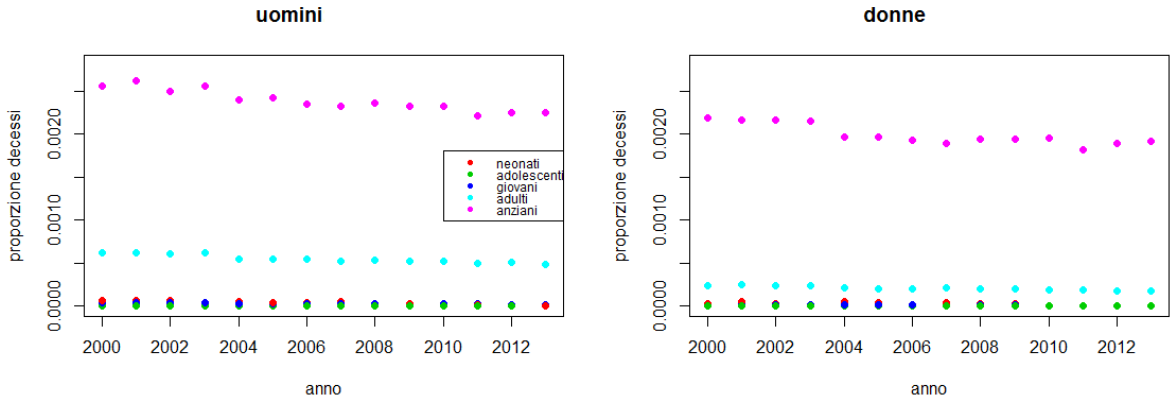

Figura 2.15: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 12: malattie dei sistema digerente

Per le malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico, considerate nel grafico 2.16, invece, la proporzione di donne decedute per queste cause supera di gran lunga quella degli uomini nella categoria delle persone con più di 70 anni. Nelle prime si può però vedere un certo decremento nel corso del tempo, mentre negli uomini l'andamento è pressochè costante, infatti negli ultimi anni di osservazione i valori per i due sessi sono quasi analoghi, malgrado la sensibile differenza iniziale.

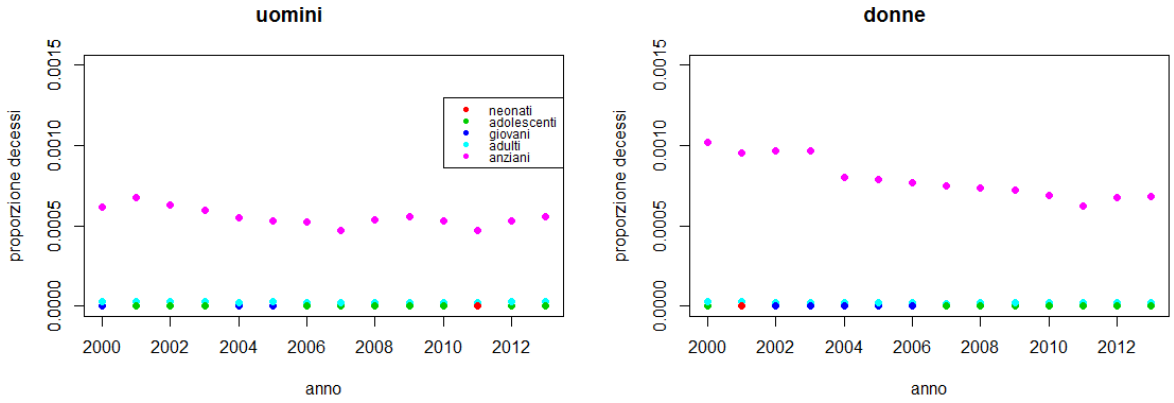

Figura 2.16: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 13: malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico

La causa relativa a malattie del sistema genitourinario e complicanze legate alle gravidanza, vede ancora una volta il sesso maschile con valori più elevati rispetto a quello femminile, come riportato nel grafico 2.17. Ancora una volta la categoria con valori maggiori è quella relativa agli anziani, questo fa dedurre che le complicanze della gravidanza non incidano quasi per nulla tra le cause di morte, se così fosse infatti si noterebbero valori più alti nelle donne appartenenti alla categoria delle "giovani" e delle "adulte", piuttosto che per le "anziane".

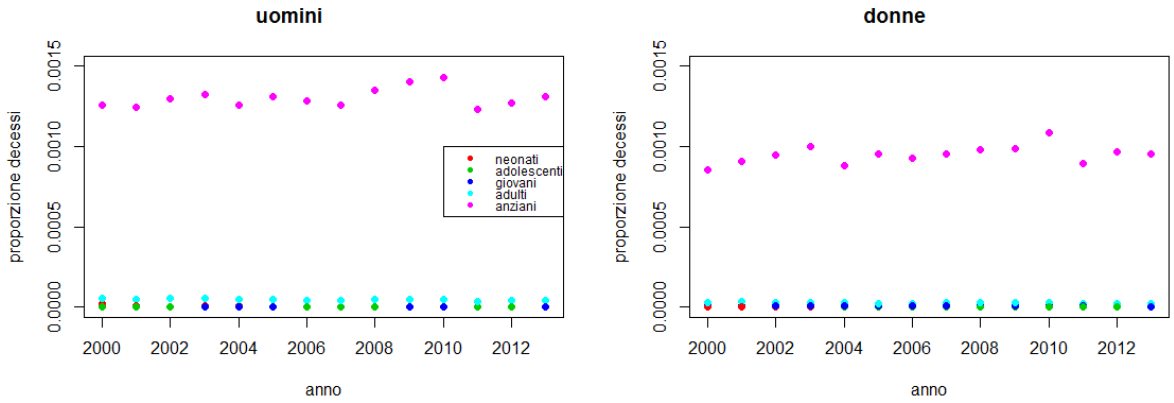

Figura 2.17: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 14: malattie del sistema genitourinario e complicanze legate alla gravidanza

Nel grafico 2.18 si configura una situazione completamente diversa per quanto attiene alle categorie di età rispetto a tutti gli altri casi, infatti, come è possibile

vedere, l'unica classe ad essere colpita è quella dei neonati, con valori leggermente superiori per i maschi, questo è da ricercarsi nella natura stessa della causa, si tratta infatti di deceduti per condizioni originarie del periodo perinatale e anomalie o malformazioni congenite.

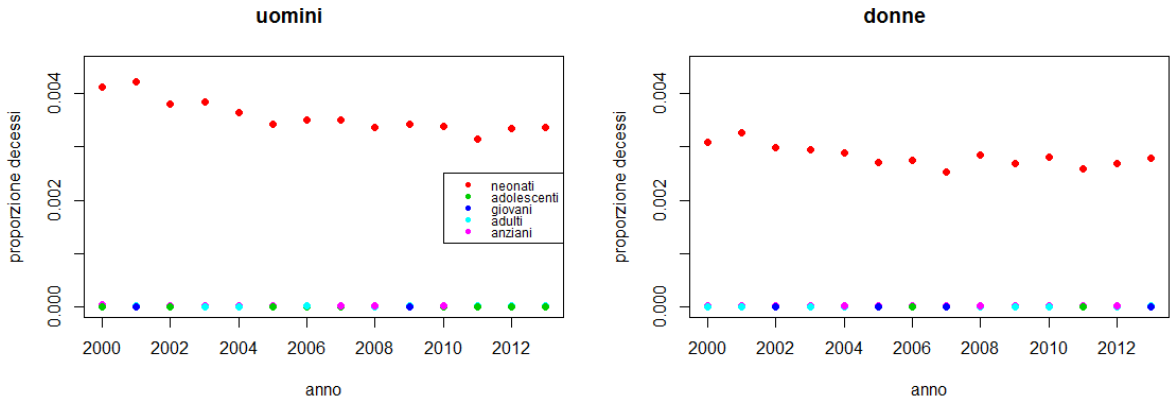

Figura 2.18: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 15: condizioni originarie del periodo perinatale e anomalie o malformazioni congenite

Per le cause esterne, riportate nel grafico 2.19, si può notare come per gli uomini non ci siano valori elevati delle proporzioni solo per le persone anziane, ma anche per gli adulti e i giovani, a differenza delle donne dove i valori riferiti a queste categorie sono comunque piuttosto bassi. Anche in questo caso tali conclusioni sono scontate se si considera la natura della causa di morte; infatti ricercando con maggior dettaglio le cause agglomerate in questa macro-categoria dalla *Intermediate List* si può riscontrare come ne facciano parte tutta una serie di incidenti come quelli stradali, subacquei, annegamenti, avvelenamenti da fumo, alcool, sostanze stupefacenti e suicidi, che colpiscono molto spesso persone attive lavorativamente o nel tempo libero, quindi non solo le persone anziane.

Per altre malattie non comprese o ignote, presenti nel grafico 2.20, l'andamento relativo alle persone anziane tra uomini e donne è molto simile, per le altre categorie invece si possono notare valori leggermente più elevati per gli adulti e i neonati negli uomini, mentre per le donne non vi è la presenza di questo andamento.

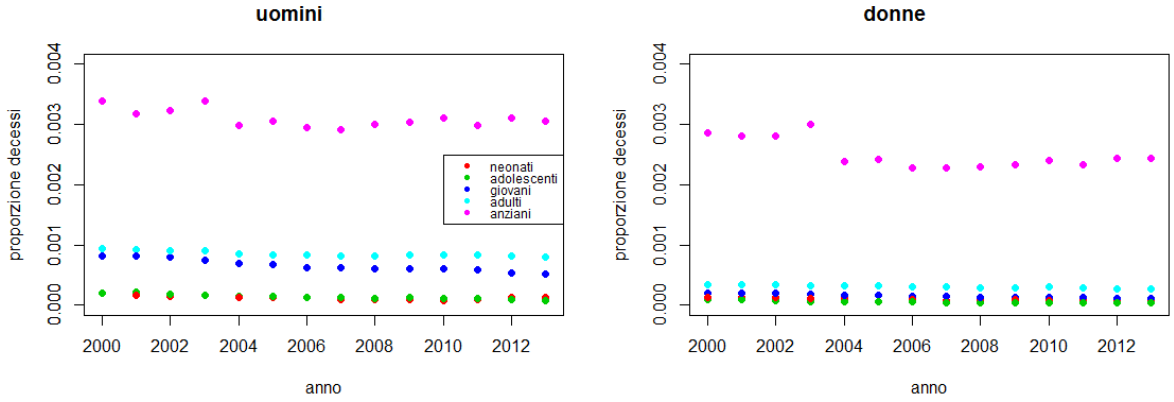

Figura 2.19: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 16: cause esterne

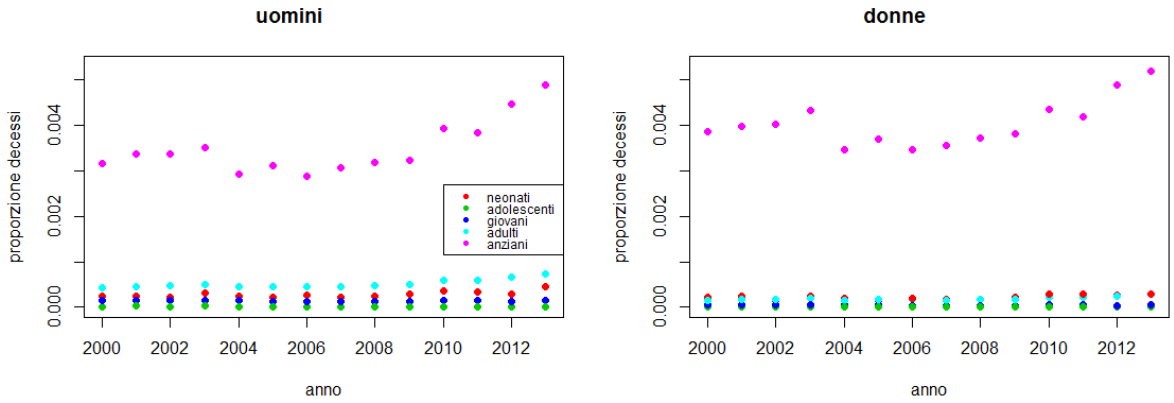

Figura 2.20: proporzione dei decessi per sesso, anno di riferimento e classe di età, causa 17: altre cause non comprese o cause ignote

Capitolo 3

I modelli

Per le analisi condotte, con i dati descritti nel capitolo precedente, sono state usate due classi di modelli, vale a dire i Modelli Lineri Generalizzati (GLM) con risposta di tipo Poisson e Binomiale Negativa e i modelli ad Inflazione di Zeri, con le stesse distribuzioni (ZIP per il modello di Poisson e ZINB per la Binomiale Negativa). Nel resto del capitolo saranno approfonditi questi modelli, con maggior dettaglio, sul profilo teorico. Mentre nel capitolo successivo si potranno vedere i risultati che sono stati ottenuti applicandoli ai dati relativi alle cause di morte nella popolazione francese; verrà poi presentato un confronto per poter individuare quale tra quelli considerati sia il modello che meglio illustra i dati analizzati.

Si tratta di modelli per dati di tasso, infatti viene usata come variabile risposta la numerosità di individui morti e tra le variabili esplicative viene inserita la popolazione di riferimento con *offset*, il cui coefficiente associato sarà quindi posto pari a 1. Questo uso della variabile relativa alla popolazione a rischio permette di definire la dimensione del contesto, essa non risulta particolarmente interessante da un punto di vista interpretativo, ma risulta utile per considerare i tassi di mortalità per causa, permette infatti di tener conto della differenza che esiste nelle popolazioni degli esposti al rischio, giacchè queste possono essere anche di molto differenti. La variabile relativa alla popolazione viene anch'essa trasformata attraverso il logaritmo (come accade per la variabile risposta, attraverso il legame canonico). Se si definisce con μ_i il valore atteso per la generica osservazione i-esima, β_{ij} il parametro relativo alla j-esima variabile dell'i-esima osservazione e con P_i la popolazione relativa alla

generica i-esima osservazione, si ottiene:

$$\log \mu_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \lambda \cdot \log P_i$$
$$\log \mu_i - \lambda \cdot \log P_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ...$$
$$\log \frac{\mu_i}{P_i} = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ...$$

Si ricorda che il parametro relativo alla popolazione, quindi in questo caso λ, viene di conseguenza posto pari a 1.

Questa specificazione permette di modellare il tasso relativo alla i-esima osservazione, ossia il rapporto tra i morti con una certa combinazione di caratteristiche e la popolazione relativa, rappresentato da: $\frac{\mu_i}{P_i}$.

3.1 I GLM

I Modelli Lineari Generalizzati (GLM) sono stati introdotti nel 1972, da Nelder e Wedderburn [Nelder e Wedderburn 1972] con lo scopo di modellare una serie di situazioni in cui i dati possono essere sia distribuiti normalmente che non.

Nel caso considerato la variabile risposta è un conteggio, quindi di tipo discreto, l'utilizzo di modelli lineari risulta forzato in quanto la distribuzione normale ha dominio sull'intero asse reale mentre quello della variabile considerata è l'insieme dei numeri naturali (positivi e interi).

Nei GLM non vi è la necessità che la variabile risposta appartenga ad una distribuzione normale, in particolare basta che questa appartenga ad una famiglia di distribuzione esponenziale; affinchè ciò avvenga è sufficiente che sia possibile scrivere la funzione di probabilità nella forma [Azzalini 2001]:

$$f(y;\theta,\psi) = exp\left\{\frac{(y \cdot \theta - b(\theta))}{a(\psi)} + c(y,\psi)\right\}$$

dove:

- θ è detto "parametro naturale" ed è ignoto;
- ψ è detto "parametro di scala" ed è anch'esso generalmente ignoto;
- $a(\psi) = \frac{\psi}{w}$, spresso w=1 quindi $a(\psi) = \psi$.

La particolare distribuzione viene specificata attraverso diverse funzioni $a(\cdot), b(\cdot)$ e $c(\cdot)$ che sono conosciute. Deve essere inoltre verificato che la variabile risposta Y abbia dominio indipendente dai parametri naturale e di scala.

Se queste condizioni sono verificate, si può dire che Y è una variabile casuale che appartiene alla famiglia di dispersione esponenziale, di parametro θ e si scrive: $Y \sim EF\left(b(\theta), \frac{\psi}{w}\right)$.

I GLM sono modelli più flessibili rispetto al modello lineare in cui la variabile risposta viene trasformata attraverso una generica funzione $g(\cdot)$. I Modelli Lineari Generalizzati sono caratterizzati dalla presenza di tre componenti:

- una componente casuale: la variabile Y_i deve appartenere ad una famiglia di distribuzioni che dipenda dal parametro naturale θ e di media μ_i;
- una componente sistematica: rappresentata dal predittore lineare che indica la relazione tra le variabili esplicative e i coefficienti $\eta_i = \sum_{j=1}^{p} x_{ij} \cdot \beta_j$;
- una funzione legame: $g(\mu_i) = \eta_i$ per mettere in relazione valore medio e predittore lineare.

3.1.1 La stima dei parametri nei GLM

La stima del vettore dei parametri β che rappresentano la relazione intercorrente tra le variabili esplicative x_i e il valore atteso μ_i, viene fatta attraverso la massimizzazione della verosimiglianza[Azzalini 2001].

Grazie all'ipotesi di indipendenza delle componenti è possibile scrivere la log-verosimiglianza, ossia il logaritmo della verosimiglianza, come:

$$l(\beta) = \sum_{i=1}^{n} \frac{y_i \cdot \theta_i - b(\theta_i)}{a_i(\psi)} + c(y_i, \psi)$$

Se il parametro ψ risulta ignoto [1] è sufficiente considerarlo come parametro di disturbo e condizionarsi ad esso per l'inferenza; questo non inciderà sulla stima di

[1] Per la stima del parametro di dispersione ψ è possibile intraprendere due strade:

1. attraverso la stima di massima verosimiglianza: avendo a disposizione la stima del vettore $\widehat{\beta}$ è possibile calcolare quella dei valori medi stimati $\widehat{\mu}$ e quindi ottenere $\widetilde{\psi}$;
2. sfruttare la relazione: $Var(Y_i) = E[(Y_i - \mu_i)^2] = a_i(\psi)V(\mu_i)$ per ottenere una stima più robusta della precedente e più stabile:

β essendo ψ e θ ortogonali tra loro. Le equazioni di verosimiglianza: $l'(\theta) = \dfrac{\partial l(\theta)}{\partial \theta}$, che si ottengono derivando la log-verosimiglianza e ponendo tale derivata pari a 0, spesso non sono risolvibili in forma esplicita, a prescindere dalla conoscenza o meno di ψ.

Per risolvere queste equazioni, qualora non ammettano soluzioni esplicite, è necessario ricorrere all'uso di un algoritmo iterativo basato sulla risoluzione di una successione di problemi di stima di minimi quadrati: l'algoritmo *scoring di Fisher*.

Questo algoritmo rappresenta una variante del metodo di Newton-Raphson, il quale si basa su un'approssimazione in serie di Taylor del primo ordine. Rispetto al metodo di Newton-Raphson, al posto della matrice Hessiana, come derivata seconda, nello *scoring di Fisher* si usa la matrice di informazione attesa, cambiata di segno, calcolata per il generico $\beta_{jt}^{(t)}$; si ottiene, quindi, per la stima di β al passo t+1:

$$\beta^{(t+1)} = \beta^{(t)} + I(\beta^{(t)})^{(-1)} l'(\beta^{(t)})$$

con $I(\beta^{(t)})_{jk} = E\left(-\dfrac{\partial^2 l}{\partial\beta_j \partial\beta_k}\right)$ per $j, k = 1, 2, 3, ..., p$ che in termini di matrici si scrive $I(\beta) = X^T \widetilde{W} X$. Dopo opportuni calcoli e riportando la dicitura matriciale, si ottiene che l'iterazione *scoring* al passo t+1[2] è rappresentata da:

$$\beta^{(t+1)} = (X^T \widetilde{W}^{(t)} X)^{-1} X^T \widetilde{W}^{(t)} z^{(t)}$$

Dove:

- X è la matrice delle variabili esplicative;

$$a_i(\widehat{\psi}) = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

che spesso coincide con ψ_i. Nel caso di modello normale con funzione legame identità $\psi = s^2$, è quindi possibile sfruttare tale relazione intercorrente.

[2]Ad ogni passo l'algoritmo procede ad una stima ai minimi quadrati ponderati, grazie alla presenza di $\widetilde{w}_i$, da cui il nome di "algoritmo dei minimi quadrati pesati iterati" (IWLS), con pesi e valori di z_i che cambiano ad ogni passo. Il suddetto algoritmo è composto da due passi principali:

1. avendo $\beta^{(t)}$ è possibile calcolare $z^{(t)}$ e $\widetilde{W}^{(t)}$;
2. grazie alle quantità ottenute al passo precedente è possibile stabilire $\beta^{(t+1)}$.

Come valori inizali, per agevolare la procedura del calcolo dell'algoritmo, è possibile porre $z_i^{(0)} = g(y_i)$ e $\widetilde{W}_i^{(0)}$ pari alla matrice identità.

- $\widetilde{W}$ è una matrice diagonale, il cui l'i-esimo elemento è dato da:

$$\widetilde{w}_i = \frac{1}{var\{Y_i\}} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{w_i}{\psi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2;$$

- $z^{(t)}$ rappresenta un'approssimazione attraverso lo sviluppo in serie di Taylor del primo ordine del vettore $g(y)$ in funzione di $g(\mu_i)$ che descrive il predittore lineare:

$$g(y_i) \cong g(\mu_i) + (y_i - \mu_i) \cdot g'(\mu_i) = \eta_i + (y_i - \mu_i) \cdot \frac{\partial \eta_i}{\partial \mu_i}.$$

Sono valide, anche in un contesto di modelli lineari generalizzati, le usuali proprietà asintotiche degli stimatori di massima verosimiglianza: la distribuzione approssimata dello stimatore di massima verosimiglianza è una Normale $\widehat{\beta} \sim N_p(\beta, I(\beta)^{-1})$ con $I(\beta) = X^T \widetilde{W} X$, e viene stimata anch'essa con la massima verosimiglianza.

Anche test ed intervalli di confidenza, con ψ noto, sono quelli tipici:

$$\text{con } H_0 : \beta_j = 0 \text{ la statistica è } Z = \frac{\widehat{\beta}_j}{\sqrt{\{(X^T \widetilde{W} X)^{-1}_{jj}\}}} \sim N(0,1)$$

3.1.2 Le analisi diagnostiche nei GLM e il confronto tra modelli

Uno degli obiettivi che si vuole raggiungere con la creazione di un modello è quello di far si che si riesca a spiegare la variabile risposta attraverso una serie di variabili esplicative, raggiungendo un buon adattamento del suddetto ai dati di cui si dispone. Fondamentale è che il modello sia il più semplice possibile, ossia che rappresenti in modo efficace ma meno complesso il fenomeno di interesse che si sta studiando [Azzalini e Scarpa 2012]. Per poter considerare questi punti si utilizza il confronto tra il modello corrente e quello saturo, ossia quello contenente tanti parametri quante sono le osservazioni (inutile dal punto di vista interpretativo perchè non semplifica il fenomeno di interesse ma utile per il confronto con il modello corrente). Nel primo modello i parametri saranno indicati con $\widehat{\theta}_i$, mentre nel secondo, dove corrispondono alle osservazioni, con $\widetilde{\theta}_i$.

Il metodo principe per il confronto tra un modello e quello saturo è la Devianza, in particolare, definendo come in precedenza $l(\cdot)$ la log-verosimiglianza, si ottiene [Azzalini 2001]:

$$W(y) = -2[l(\widehat{\beta}) - l(\widetilde{\beta})] = -2\sum_i \frac{w_i}{\psi}[(y_i\widehat{\theta}_i - b(\widehat{\theta}_i)) - (y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i))] = \frac{\sum_i d_i}{\psi}$$

che prende il nome di Devianza normalizzata, questa si distribuisce come un χ^2_{n-p} (dove p è il numero di parametri stimati)[3], mentre $D(y;\widehat{\mu}) = \sum_i d_i$ è la Devianza e il generico elemento d_i è il contributo dell'i-esima osservazione.

Si può fare il confronto, anzichè tra il modello corrente e il modello saturo, anche tra modelli annidati, ossia modelli che hanno un'analoga specificazione nei quali il modello più grande, M_1, si differenzia dal più piccolo, M_2, solo per la presenza di alcuni parametri aggiuntivi del predittore lineare e per null'altro ($M_2 \subset M_1$).

La funzione di devianza normalizzata, dove p_1 e p_2 sono rispettivamente il numero di parametri dei due modelli considerati ($p_1 > p_2$), prende la forma:

$$\frac{D(y;\widehat{\mu_2}) - D(y;\widehat{\mu_1})}{\psi} \xrightarrow{d} \chi^2_{p1-p2}$$

Per studiare la bontà di adattamento del modello ai dati, in particolare da un punto di vista grafico, è possibile calcolare diversi tipi di residui; nei modelli lineari generalizzati, a differenza di quello lineare, non esiste una distinzione della variabile risposta tra la componente erratica e quella sistematica. Per questo ultimo motivo sono state elaborate più definizioni di residui:

1. residui di risposta: rappresentano una traslazione della variabile dipendente, non sono molto usati perchè ereditano l'eteroschedasticità della variabile risposta; non aiutano quindi nelle analisi diagnostiche:

$$r_i = y_i - \widehat{\mu}_i \quad \text{con} \quad \widehat{\mu}_i = g^{-1}(\widetilde{X_i^T}\beta);$$

2. residui di Pearson: sono i residui di risposta riscalati per la $Var(\mu_i)$ (che contiene l'eteroschedasticità), si tratta di un'estensione diretta del residuo standardizzato nel modello linerae:

$$r_{iP} = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}} \quad \text{con} \quad Var(Y_i) = a(\psi)V(\mu_i)$$

[3]Nel caso di ψ non noto, invece, non è più assicurato che la Devianza normalizzata si distribuisca come un χ^2, in ogni caso, anche se viene comunque usata questa distribuzione, il grado di approssimazione peggiora notevolmente.

3. residui di devianza: derivano dall'estenzione del concetto di residuo, come contributo individuale alla devianza, nel modello lineare:

$$r_{iD} = sgn(Y_i - \widehat{\mu}_i)\sqrt{d_i} \text{ dove } d_i \text{ è il contributo individuale alla devianza e } \sum_{i=1}^{n} r_{iD}^2 = D$$

4. residui di lavoro o *working residuals*: utilizzano l'algoritmo *scoring* di Fisher:

$$r_i = z_i - \eta_i \text{ dove } z_i \text{ rappresenta il primo passo dell'approssimazione di } g(y_i) \text{ per Newton-Raphson } z_i = \eta_i + (y_i - \widehat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}$$

5. residui di Anscombe: variante dei residui di Pearson, che grazie ad una trasformata, risultano più vicini alla distribuzione normale.

3.1.3 Il GLM con risposta Poisson

Se la variabile risposta, Y, è assimilabile ad una distribuzione di tipo Poisson di parametro λ: $Y \sim P(\lambda)$, la funzione di probabilità del modello generalizzato sarà [Azzalini 2001]:

$$f(y;\lambda) = \frac{e^{-\lambda} \cdot \lambda^y}{y!} = exp\{y \cdot \log \lambda - \lambda - \log y!\}$$

Le componenti saranno rispettivamente:

- $\theta = \log \lambda, \quad \psi = 1;$
- $b(\theta) = \lambda = e^{\theta};$
- $c(Y, \psi) = -\log Y!$

Si ottiene quindi che Y appartiene alla famiglia esponenziale e si indica con $Y \sim EF(e^{\theta}, 1)$, il valore atteso è $E(Y) = e^{\theta} = \lambda$ e corrisponde anche alla varianza $Var(Y) = e^{\theta} = \lambda$. Per la scelta della funzione legame, rappresentata da $g(\cdot)$, è sufficiente che questa sia monotona e derivabile; è importante, anche se non fondamentale, che questa linearizzi e mantenga la sensatezza dei risultati. Nel caso della

distribuzione di Poisson la funzione di legame canonico[4], che è quella usata per le elaborazioni successive, è descritta dal logaritmo, ossia $\eta_i = g(\mu_i) = \log \mu_i$, ma è possibile avvalersene anche di altre, come la funzione identità; la funzione legame ha lo scopo di controllare l'eteroschedasticità dei dati.

Il modello lineare generalizzato con risposta di Poisson può essere usato in diversi contesti: oltre a modellare variabili di tipo conteggio, tale distribuzione è anche la distribuzione limite della Binomiale, quando il numero di prove binarie è molto elevato, le probabilità sono basse e il *dataset* è numeroso (situazione analoga a quella relativa ai dati sulle cause di mortalità per la popolazione francese), in questo contesto è più flessibile del secondo in quanto prevede la possibilità che si verifichino più eventi. Può adattarsi a tabelle di contingenza e di frequenza, oppure può essere sfruttato, come viene usato in questo elaborato, per situazioni in cui si intende modellare il tasso di un certo evento, introducendo una variabile relativa al dominio nel quale è contestualizzato il fenomeno. Questa distribuzione presenta però un importante limite, ossia la media e la varianza sono assunte uguali, in situazioni in cui non si verifica questo vi è la necessità di provvedere con altre soluzioni.

Per quanto riguarda la Devianza questa coincide con quella normalizzata, essendo $\psi = 1$, e vale:

$$D = -2\sum\{(y_i \log \widehat{\mu}_i - \widehat{\mu}_i) - (y_i \log y_i - y_i)\} = 2\sum_i \left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i} - y_i + \widehat{\mu}_i\right)\right\}$$

Ponendo per convenzione $0 \cdot \log 0 = 0$, se si verifica che $\sum_i y_i = \sum_i \widehat{\mu}_i$ (ossia in presenza di legame canonico), allora la precedente espressione si semplifica notevolmente e diventa:

$$D = 2\sum_i y_i \log \frac{y_i}{\widehat{\mu}_i}$$

[4]Dalla letteratura in materia, si evince che usare come funzione legame quella canonica ha diversi vantaggi, quindi quando questo è possibile risulta conveniente operare tale scelta. In particolare, si riconoscono benefici per quanto riguarda:

- la riduzione dell'informazione: se ψ è noto vale il teorema di fattorizzazione, infatti risulta che $\sum_i y_i \cdot \widetilde{x}_i^T$ è statistica sufficiente minimale per β; se ψ non è noto ne è comunque parte;
- le equazioni di verosimiglianza risultano semplificate, come si verdrà a breve con la Devianza;
- la matrice di informazione attesa coinciderà con la matrice di informazione osservata, questo perchè la prima non dipenderà più da Y che risulta l'unico elemento aleatorio.

3.1.4 Il GLM con risposta Binomiale Negativa

La distribuzione Binomiale Negativa è un'estensione della distribuzione di Poisson e, a differenza di quest'ultima, permette di modellare situazioni in cui la varianza è maggiore della media.

La funzione di probabilità, se la variabile $Y \sim NegBin(k, p)$ è rappresentata da *Generalized linear models*:

$$P(Y = y) = \frac{(y-1)!}{(y-k)!(k-1)!}\rho^k(1-\rho)^{y-k} \quad \text{con} \quad y = k, k+1, k+2, ...$$

Dove:

- y rappresenta il numero di prove necessarie per ottenere k successi;
- k indica una sorta di tempo di attesa, per avere quel certo numero di successi;
- ρ è la probabilità di successo per la singola prova.

Non è disponibile, con k ignoto, una forma esplicita della funzione di probabilità scritta in modo da rendere evidente l'appartenenza di questa distribuzione alla famiglia di dispersione esponenziale, ma si può ottenere come estensione della distribuzione di Poisson, modellata con una distribuzione Gamma[5] [Usai 2011].

Con k noto invece è possibile ricavare:

$$f(y;\rho) = exp\left\{log\binom{y-1}{k-1} + k \cdot log\frac{\rho}{1-\rho} + y \cdot log(1-\rho)\right\}$$

i cui fattori espressi come componenti della famiglia di dispersione esponenziale sono:

- $\theta = log(1-\rho), \quad \psi = 1;$
- $b(\theta) = k \cdot log\left(\frac{\rho}{1-\rho}\right);$
- $c(y, \psi) = log\binom{y-1}{k-1}, \quad$ che rappresenta il coefficiente binomiale.

Il legame canonico per questa distribuzione, che è quello che verrà usato anche nelle elaborazioni successive (i cui aspetti migliorativi sono già stati trattati nella

[5]Si ricorda, in particolare, che una distribuzione Gamma di parametri κ: $Y \sim Gamma(\kappa, \kappa)$, quando κ è intero coincide con una distribuzione Binomiale Negativa di parametri κ e π: $Y \sim NegBin(\kappa, \pi)$ con $\pi = 1/(1 + \delta/\kappa)$ [Pace e Salvan 2001]

sezione precedente per il GLM di Poisson, ma che valgono per tutte le distribuzioni), è rappresentato da $log\frac{\mu}{\mu+k}$.

Per quanto riguarda la devianza, questa equivale a:

$$D = 2\sum_i \left[y_i \cdot log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (k + y_i) \cdot log\left(\frac{k+\widehat{\mu}_i}{k+y_i}\right)\right]$$

3.2 I modelli *Zero Inflated*

I modelli ad inflazione di zeri non sono dei modelli lineari generalizzati ma ne rappresentano una loro estensione. Sono idonei a spiegare quelle situazioni nelle quali c'è una massiccia presenza di valori assunti pari a zero dalla variabile risposta; in particolare, l'incidenza di questi valori è superiore rispetto a quella che ci si può aspettare dalla distribuzione specifica.

L'idea che sta alla base di questa inefficienza, causata da una sovradispersione, è quella che il processo generatore dei dati che produce gli zeri sia diverso rispetto a quello che produce valori superiori a zero.

Vengono quindi usati modelli nati dalla manifestazione congiunta di due fenomeni non singolarmente osservabili, si tratta di due processi indipendenti. Se, quindi, la variabile risposta, legata al fenomeno oggetto di interesse, è rappresentata da Y, questa viene scissa in due parti: $Y = W \cdot Z$. La distribuzione W è legata al verificarsi dell'evento $Y = 0$ mentre Z all'evento $Y = y > 0$, con $y \in \mathbb{N}$.

I modelli che sono utilizzati in questo elaborato sono rispettivamente l'estensione della distribuzione di Poisson (ZIP) e quella relativa alla Binomiale Negativa (ZINB); che sono le uniche[6] che prevedono questa possibile estensione (è necessario infatti una distribuzione che si adatti a conteggi).

[6]Esistono delle ulteriori estensioni che danno luogo ai modelli di tipo *Hurdle* ossia gli *zero altered model* o *two part model*, che però riguardano comunque solo le distribuzioni di Poisson e Binomiale Negativa. Questi modelli, che non verranno trattati in questo elaborato, presuppongono la presenza di due popolazioni veramente distinte: una per tutte le osservazioni con valore pari a 0, che rappresenterebbero gli "zero strutturali", e l'altra che prevede valori 0, questi però sono gli "zero campionari", o valori maggiori di zero. [Viviano 2008]

3.2.1 Il modello ZIP

Se una varibile ha una distribuzione di tipo Poisson, ci si aspetta che il numero di volte in cui non si verifica alcun evento, ossia in cui si presenta uno 0 in n dati, sia circa pari a $n \cdot \mathbb{P}(y_i = 0) = n \cdot e^{-\lambda}$. Quando invece questa quantità raggiunge proporzioni più elevate non è possibile usare questo tipo di distribuzione per modellare le osservazioni.

Nello *Zero Inflated Poisson* vi è la presenza di una variabile Y che nasce dalla manifestazione congiunta di $Y = W \cdot Z$: nella fattispecie la variabile W assume una distribuzione di tipo Bernoulliano con parametro $(1-\pi)$, mentre Z una distribuzione di tipo Poisson con parametro λ.

La variabile Y diventa la risposta di un modello di regressione congiunto, specificato per le due componenti:

$$\begin{cases} log(\lambda) = \widetilde{x}_{(1)}^T \cdot \beta & \text{rappresenta la regressione di Poisson;} \\ logit(1-\pi) = log\left(\dfrac{\pi}{1-\pi}\right) = \widetilde{x}_{(2)}^T \cdot \gamma & \text{parte logistica relativa alla Bernoulli.} \end{cases}$$

Dove le $\widetilde{x}_{(1)}^T$ e $\widetilde{x}_{(2)}^T$ rappresentano le covariate, mentre β e γ sono i parametri stimati nelle due regressioni, di ciascuna delle due componenti modellate come due GLM.

La $\mathbb{P}(Y = y)$ può quindi essere scomposta in due parti:

- $\underline{y > 0} \quad \longrightarrow \quad \mathbb{P}(Y = y) = \mathbb{P}(W \cdot Z = w \cdot z) = \mathbb{P}(W = 1 \cap Z = y) = \\ = \mathbb{P}(W = 1) \cdot \mathbb{P}(Z = y) = (1 - \pi)\dfrac{e^{-\lambda} \cdot \lambda^y}{y!}$
- $\underline{y = 0} \quad \longrightarrow \quad \mathbb{P}(Y = y) = \mathbb{P}(W \cdot Z = 0) = \mathbb{P}(W = 0 \cap Z = 0) + \\ + \mathbb{P}(W = 0 \cap Z > 0) + \mathbb{P}(W = 1 \cap Z = 0) = \pi + (1 + \pi) \cdot e^{-\lambda}$

Il valore atteso della variabile originale, ossia Y, è dato dal prodotto del valore atteso delle singole distribuzioni:

$$\mathbb{E}(Y) = \mathbb{E}(W \cdot Z) = (1 - \pi) \cdot \lambda = \mu$$

mentre per la varianza, dopo opportuni calcoli e semplificazioni, si ottiene:

$$Var(Y) = \mathbb{E}(Z)^2 \cdot Var(W) + \mathbb{E}(W)^2 \cdot Var(Z) + Var(W) \cdot Var(Z) = \\ (1 - \pi) \cdot \lambda \cdot (1 + \lambda - (1 - \pi) \cdot \lambda) = \mu \cdot (1 + \pi\lambda)$$

Le covariate che influiscono sulle due parti, quella relativa al verificarsi di $Y = 0$ e quella relativa a $Y > 0$ possono essere le medesime o anche differire notevolmente, infatti i due processi sono considerati del tutto indipendenti. Spesso si tende ad usare lo stesso set di variabili esplicative al fine di vedere il diverso ruolo che ha ciascuna nei due stadi del modello. Tuttavia, malgrado vi sia una netta distinzione tra le due sottodistribuzioni, il modello deve essere applicato contemporaneamente ai dati, attraverso la stima di massima verosimiglianza.

3.2.2 Il modello ZINB

Nel modello *Zero Inflated Negative Binomial* la scomposizione della variabile dipendente in $Y = W \cdot Z$ avviene come nel modello ZIP con l'unica differenza che al posto di una distribuzione di tipo Poisson vi è la presenza di una distribuzione di tipo Binomiale Negativa: $Z \sim BiNeg(\kappa, \rho)$ (di parametri (κ, ρ) con $\rho = \delta/\kappa$ e δ indica il valore atteso) per la parte relativa a $Y > 0$. In particolare la $\mathbb{P}(Y = y)$:

- $\underline{y > 0} \quad \longrightarrow \quad \mathbb{P}(Y = y) = \mathbb{P}(W \cdot Z = w \cdot z) = \mathbb{P}(W = 1 \cap Z = y) =$
$$= \mathbb{P}(W = 1) \cdot \mathbb{P}(Z = y) = (1 - \pi)\binom{y}{k}\frac{\rho^{\kappa}}{(1-\rho)^{\kappa}}(1 - \rho)^{y}$$
- $\underline{y = 0} \quad \longrightarrow \quad \mathbb{P}(Y = y) = \mathbb{P}(W \cdot Z = 0) = \mathbb{P}(W = 0 \cap Z = 0) +$
$$+\mathbb{P}(W = 0 \cap Z > 0) + \mathbb{P}(W = 1 \cap Z = 0) = \pi + (1 - \pi) \cdot \rho^{\kappa}$$

Il valore atteso della distribuzione originale è dato da:

$$\mathbb{E}(Y) = \mathbb{E}(W \cdot Z) = (1 - \pi) \cdot \frac{\kappa}{\rho} = \mu$$

mentre per la varianza, dopo opportuni calcoli e semplificazioni, è:

$$Var(Y) = \mathbb{E}(Z)^2 \cdot Var(W) + \mathbb{E}(W)^2 \cdot Var(Z) + Var(W) \cdot Var(Z) =$$
$$\frac{\kappa}{\rho^2} \cdot (1 - \rho - \pi + \rho \cdot \pi) = \frac{k}{\rho^2} \cdot (1 - \pi) \cdot (1 - p) = \mu \cdot \frac{1 - \rho + \kappa \cdot \pi}{\rho}$$

Anche in questa estensione le covariate che influiscono sulle due parti della scomposizione potrebbero coincidere, per identificare il ruolo diverso che esse possono avere, o essere differenti, come già spiegato nella sezione relativa ai modelli ZIP. Per quanto riguarda la stima dei parametri, come in precedenza, è necessaria una procedura unica, malgrado vi sia distinzione tra le due parti, per massimizzare la verosimiglianza.

Capitolo 4

Confronto tra modelli

In questo capitolo verranno implementati ai dati i modelli descritti nel precedente, in particolare nella prima sezione si vedranno i modelli Lineari Generalizzati, mentre nella seconda i modelli *Zero Inflated*. Verrà presentato un solo modello per tipologia, scelto attraverso il test ANOVA, ossia il confronto tra le devianze di modelli annidati, presentato nella sezione relativa alle analisi diagnostiche del capitolo precedente. Nel capitolo sono presentati solo gli effetti principali e alcune caratteristiche particolarmente interessanti, mentre maggiori dettagli sono presenti nell'appendice dell'elaborato.

Come già anticipato i modelli sono distinti per sesso, in modo da poter cogliere maggiormente l'influenza di questo fattore sulle dinamiche presenti.

Le variabile e le loro codifiche sono già state presentate nel capitolo 2, di volta in volta saranno presentati i coefficienti e la relativa interpretazione.

4.1 I modelli GLM

I modelli GLM che sono stati studiati sono rispettivamente con risposta di Poisson e Binomiale Negativo il cui predittore lineare è:

$$\eta_i = \beta_1 \cdot x_{i,1} + ... + \beta_{17} \cdot x_{i,17} + \gamma_3 \cdot x_{i,18} + ... + \gamma_{22} \cdot x_{i,37} + \delta_1 \cdot x_{i,38} + \delta_2 \cdot x_{i,39} + \\ + \alpha_{40} \cdot x_{i,40} + ... + \alpha_{395} \cdot x_{i,395} + \phi_{396} \cdot x_{i,396} + ... + \phi_{715} \cdot x_{i,715}$$

Dove i coefficienti specifici rappresentano i parametri rispettivamente per:

- β: le 17 cause;
- γ: le 20 classi di età, la categoria di riferimento è rappresentata dalla prima classe ossia coloro che sono nel primo anno di età;
- δ: l'anno di osservazione e il quadrato dello stesso;
- α: le interazioni di secondo grado tra la causa, la classe di età e l'anno;
- ϕ: le interazioni di terzo grado tra la causa, la classe di età e l'anno.

Per sviluppare tali analisi è stato usato il *software R*, di volta in volta, saranno indicati i comandi utilizzati ed eventuali librerie necessarie per l'implementazione. Per i modelli GLM è stata considerata come variabile risposta il numero di decessi mentre le covariate usate sono: la popolazione, tramite *offset*, la causa, la classe di età degli individui e l'anno di osservazione riscalato. Sone state inserite anche l'anno con potenza quadra e l'interazione tra la causa, la classe di età e l'anno dell'osservazione. La scelta della funzione legame è ricaduta nel legame canonico ossia il logaritmo, sia per il GLM con risposta di Poisson che per la Binomiale Negativa: $log(\mu_i) = \eta_i$ e si è deciso di stimare un modello senza intercetta al fine di ottenere un coefficiente per ogni causa.

Per ottenere quindi l'interpretazione dei parametri è necessario applicare la trasformata inversa della funzione legame ai parametri moltiplicati per le esplicative, in questo caso la funzione esponenziale, in modo da ottenere il valore atteso della proporzione di decessi per quella particolare combinazione di caratteristiche.

4.1.1 Il GLM con risposta di Poisson

I coefficienti degli effetti principali stimati attraverso il comando *glm{....family=poisson}* sono riportati nella tabella 4.1.

	Uomini			Donne		
	Stima	*Std Error*	*p-value*	Stima	*Std Error*	*p-value*
causa 1	-8.844	0.0771	0.000	-9.056	0.08804	0.000
causa 2	-10.06	0.1312	0.000	-10.02	0.1324	0.000
causa 3	-10.18	0.1478	0.000	-10.42	0.1767	0.000

causa 4	-9.188	0.08231	0.000	-9.283	0.08918	0.000
causa 5	-26.14	237.5	0.91238	-12.8	1.00	0.000
causa 6	-8.602	0.0248	0.000	-8.671	0.06639	0.000
causa 7	-9.09	0.08809	0.000	-9.528	0.01048	0.000
causa 8	-10.34	0.1758	0.000	-10.29	0.1928	0.000
causa 9	-12.1	0.3772	0.000	-12.85	0.5559	0.000
causa 10	-10.21	0.1435	0.000	-10.40	0.1676	0.000
causa 11	-10.39	0.1777	0.000	-10.84	0.2170	0.000
causa 12	-9.487	0.1111	0.000	-9.921	0.1387	0.000
causa 13	-12.91	0.6994	0.000	-12.61	0.5910	0.000
causa 14	-11.09	0.2693	0.000	-11.87	0.831	0.000
causa 15	-5.505	0.1327	0.000	-5.772	0.01542	0.000
causa 16	-8.640	0.06622	0.000	-8.926	0.07812	0.000
causa 17	-8.414	0.05216	0.000	-8.583	0.05913	0.000
$[1-5)$	-2.122	0.1293	0.000	-2.054	0.1448	0.000
$[5-10)$	-3.632	0.2188	0.000	-3.505	0.2251	0.000
$[10-15)$	-3.792	0.2204	0.000	-3.663	0.2477	0.000
$[15-20)$	-3.114	0.166	0.000	-3.165	0.1933	0.000
$[20-25)$	-2.786	0.1509	0.000	-3.053	0.1824	0.000
$[25-30)$	-1.942	0.1172	0.000	-2.424	0.1498	0.000
$[30-35)$	-0.8726	0.09372	0.000	-1.463	0.1162	0.000
$[35-40)$	0.07107	0.08413	0.39826	-0.8649	0.1043	0.000
$[40-45)$	0.2232	0.0249	0.00682	-0.9339	0.1035	0.000
$[45-50)$	-0.06857	0.08317	0.40972	-1.029	0.1036	0.000
$[50-55)$	-0.1075	0.0315	0.19607	-1.046	0.1031	0.000
$[55-60)$	0.1087	0.08343	0.19263	-0.6109	0.1009	0.000
$[60-65)$	0.4186	0.828	0.000	-0.1096	0.09795	0.26322
$[65-70)$	0.8791	0.08117	0.000	0.3683	0.09418	0.000
$[70-75)$	1.373	0.07999	0.000	1.033	0.09145	0.000
$[75-80)$	1.935	0.07915	0.000	1.659	0.09004	0.000
$[80-85)$	2.480	0.0792	0.000	2.228	0.08972	0.000

$[85-90)$	3.017	0.076	0.000	2.898	0.08938	0.000
$[90-95)$	3.492	0.08093	0.000	3.429	0.08949	0.000
[95 e oltre)	4.004	0.08913	0.000	3.890	0.09113	0.000
anno	-0.09746	0.01201	0.000	-0.1017	0.01373	0.000
anno^2	0.00057	3.448e-05	0.000	0.000829	3.485e-05	0.000

Tabella 4.1: *Output* parziale del GLM con risposta di Poisson relativo agli effetti principali diviso per uomini e donne (le celle colorate individuano i *p-value* superiori a 0.01, i cui parametri non sono quindi significativamente diversi da 0)

Da questa tabella si evince come i vari fattori considerati influenzino la variabile risposta, in particolare il logaritmo del rapporto tra i morti e la popolazione di riferimento. Come si può notare sia per gli uomini che per le donne i coefficienti di tutte le cause sono negativi e molto simili tra i due sessi, fatta eccezione per la causa 5, il cui coefficiente per gli uomini è assumibile pari a 0. Interessante è valutare in questo specifico caso che per gli uomini la causa 5, che si riferisce a disturbi mentali e comportamentali, non sia mai significativamente diversa da 0, nemmeno nelle interazioni con le altre variabili. Importante è notare che questa causa per gli uomini presenti degli *Standard Error* troppo elevati, sintomo che potrebbe essere presente della multicollinearità. Per le classi di età invece i coefficienti pari a 0 sono ben 4 per gli uomini, e ciò sta ad indicare che l'avere un'età compresa in questi intervalli non apporta differenze significative rispetto alla categoria di riferimento, ossia ai neonati. Per le donne invece l'unica classe di età che non si discosta da quella di riferimento è quella relativa ai [60-65) anni. I valori dei coefficienti sono abbastanza simili nei due sessi ad eccezione di alcune classi, come la [40-44) dove si ha addirittura un cambio di segno e per altre dove la differenza è anche di mezzo punto (come per le classi [25-30) e [30-35) ad esempio). Per quanto concerne l'anno di rilevazione invece questo risulta significativo per entrambi i sessi sia al primo grado che al secondo grado, nel primo caso risulta negativo mentre nel secondo è positivo.

Per una maggiore chiarezza nell'interpretazione si riportano ora un paio di possibili situazioni di cui si vuole stimare la proporzione di decessi[1]:

[1]Si considereranno come esempi solo coloro che sono deceduti nel primo anno, quindi la categoria

- uomo, causa 1 (malattie infettive), classe di età [0-1), anno 2001: $exp(-8.844 - 0.09746 + 0.00057) = 0.00013$, la proporzione di decessi con queste caratteristiche è pari allo 0.013%;
- donna, causa 15 (alcune condizioni originarie del periodo perinatale e anomalie o malformazioni congenite), classe di età [0-1), anno 2013: $exp(-5.772 - 0.1017 * 13 + 0.000829 * (13^2) + 0.0789 * 13) = 0.00266$, la proporzione di decessi con queste caratteristiche è pari a 0.27%.

I modelli presentano una Devianza Residua pari rispettivamente a 11313 per il GLM relativo agli uomini e 15258 per le donne. Tali valori sono piuttosto elevati, considerando che andrebbero confrontati con un χ^2 con un numero di gradi di libertà pari di 4283, come spiegato nel capitolo 3: sintomo che il fenomeno generatore dei dati non è correttamente specificato.

Si passa ora alle analisi grafiche dei residui per confermare se i modelli appena stimati abbiano un buon adattamento ai dati, come descritto nella sezione dedicata del capitolo 3.

Il primo grafico analizzato, il 4.1, mostra l'interpolazione tra le proporzioni osservate e quelle stimate dal modello GLM con risposta di Poisson; questi punti dovrebbero stare sulla retta bisettrice del primo e terzo quadrante. Dal grafico si può notare come vi sia una discreta situazione sia per gli uomini che per le donne, le seconde hanno una rappresentazione forse peggiore dei primi, soprattutto per valori intermedi ed elevati.

Il secondo grafico analizzato, il 4.2, riporta i residui di devianza (descritti nel capitolo 3) interpolati ai valori predetti dal modello.

Come è possibile notare, la situazione non è ben chiara, sia per gli uomini che per le donne vi sono dei valori predetti molto più piccoli rispetto agli altri. Un'approfondita indagine porta a scoprire che tutti questi valori sono imputabili, sia per le donne che per gli uomini, alle caratteristiche "causa di morte numero 5 (ossia disturbi mentali e comportamentali)" e "classe di età 2", ossia i neonati che non hanno ancora compiuto un anno di età. Dalla rimozione di queste osservazioni si ottiene il grafico 4.3, che porta ad osservare con maggior precisione le rimanenti combinazioni.

di rifermento per l'età, per una mera questione di comodità nei calcoli, non è così necessario

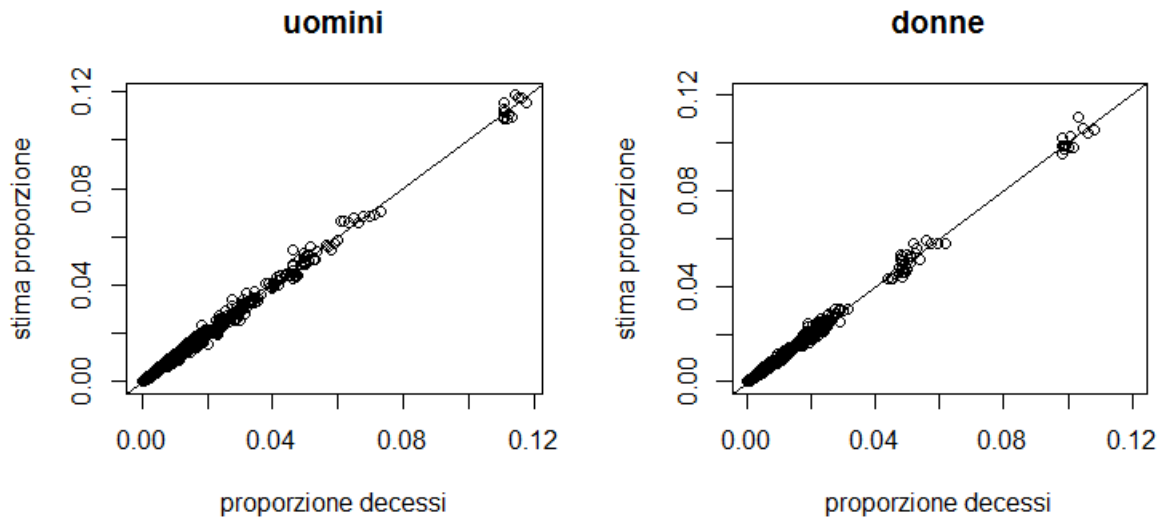

Figura 4.1: Stima della proporzione contro valore osservato, diviso per sesso: GLM Poisson

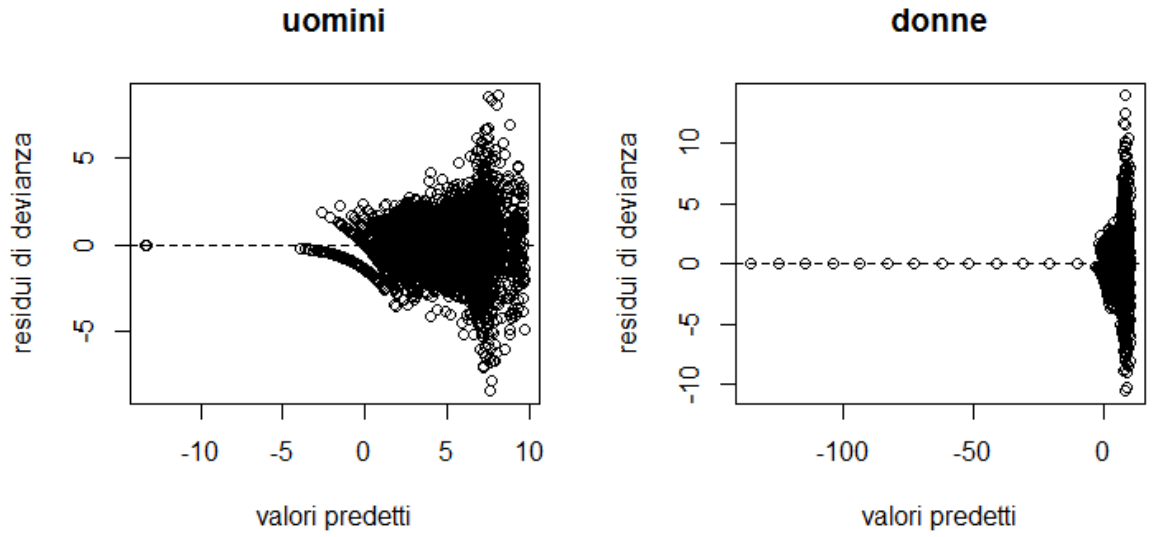

Figura 4.2: Residui di devianza contro i valori predetti diviso per sesso: GLM Poisson

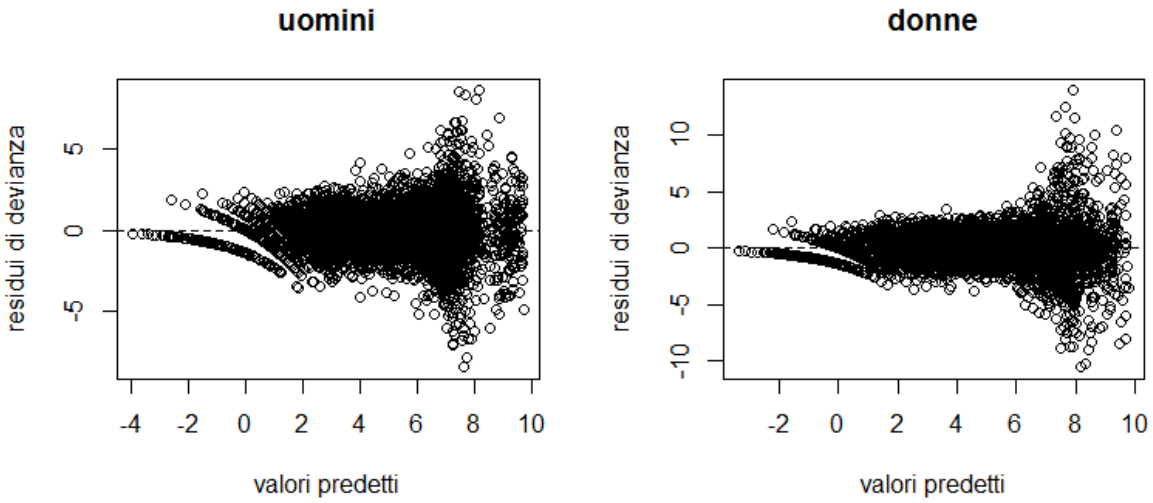

Figura 4.3: Residui di devianza contro i valori predetti diviso per sesso (senza la combinazione "causa 5 ed "età [0-1)": GLM Poisson

La situazione riportata mostra che vi è un'adeguata ripartizione tra valori positivi e negativi dei residui in entrambi i grafici, la cui media sarà attorno allo 0; si nota tuttavia anche la presenza di disomogeneità soprattutto negli estremi dei valori predetti, indice di eteroschedasticità: trattandosi di un modello GLM questa è presente per costruzione, in questo caso tale andamento potrebbe indicare che il Poisson non sia il modello adeguato a questi dati, sembra che vi sia la presenza di una situazione in cui la varianza cresce più velocemente della media: infatti nell'estremo inferiore dei valori predetti il range dei residui è molto più piccolo rispetto all'estremo superiore.

Il grafico successivo, il 4.4, riporta l'interpolazione tra i residui di lavoro e i valori predetti, questo grafico è utile per individuare l'adeguatezza della funzione legame usata. Anche in questo caso vi sono i valori relativi alla combinazione "causa 5" e "classe di età 2" che non permettono di capire in modo chiaro l'andamento del grafico, dopo la rimozione, si ottiene il grafico 4.5. La situazione anche se più nitida non è molto rassicurante, in quanto i punti dovrebbero giacere sulla bisettrice mentre, soprattutto per valori centrali, non è così.

L'ultimo grafico considerato, il 4.6, riporta i punti influenti, ossia quei punti la

introdurre tutti i parametri relativi agli effetti congiunti

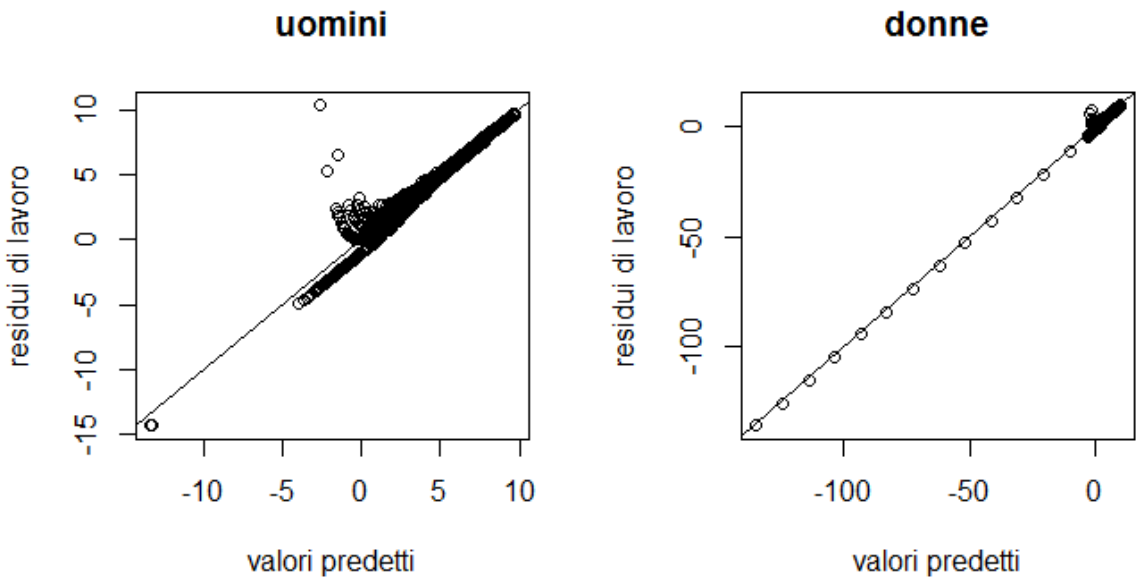

Figura 4.4: Residui di lavoro contro valori predetti diviso per sesso: GLM Poisson

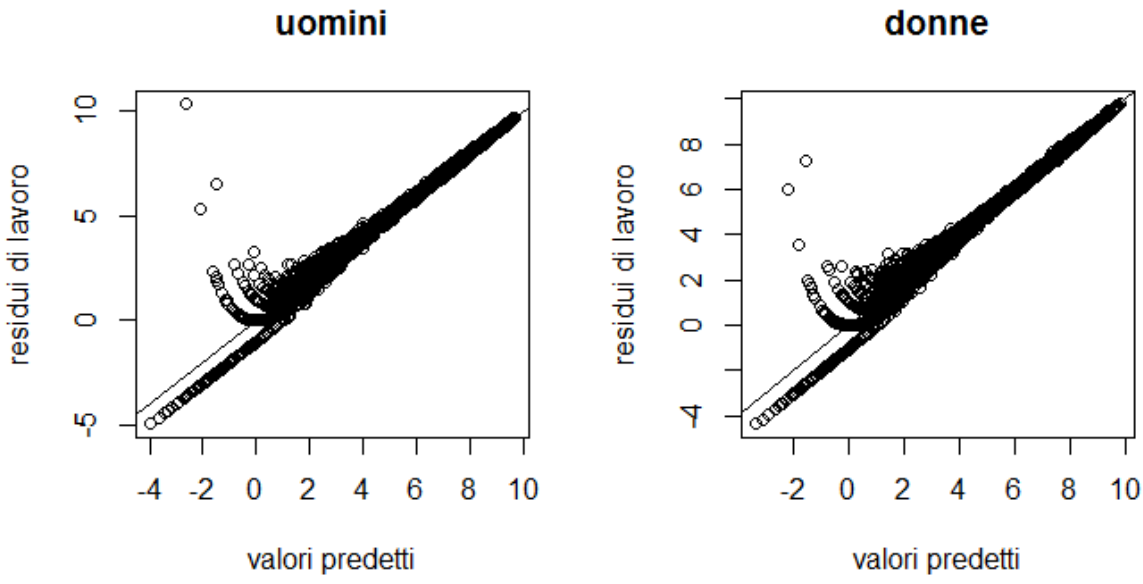

Figura 4.5: Residui di lavoro contro valori predetti diviso per sesso (senza la combinazione "causa 5" e "età [0-1)": GLM Poisson

cui presenza o assenza determina un diverso modello stimato, per uomini e donne. In dettaglio in ascissa vi è la misura della devianza standard, che porta all'individuazione di punti leva, mentre in ordinata sono presenti i residui studentizzati, l'area dei cerchi rappresenta la distanza di Cook. Più precisamente:

- punti di leva: punti con un'elevata devianza standard;
- residui studentizzati: le osservazioni che presentano un residuo studentizzato elevato (residui riscalati per la specifica devianza standard) sono chiamate *outlier* o valori anomali;
- distanza di Cook: misura dell'influenza che ha ogni valore sulle previsioni.

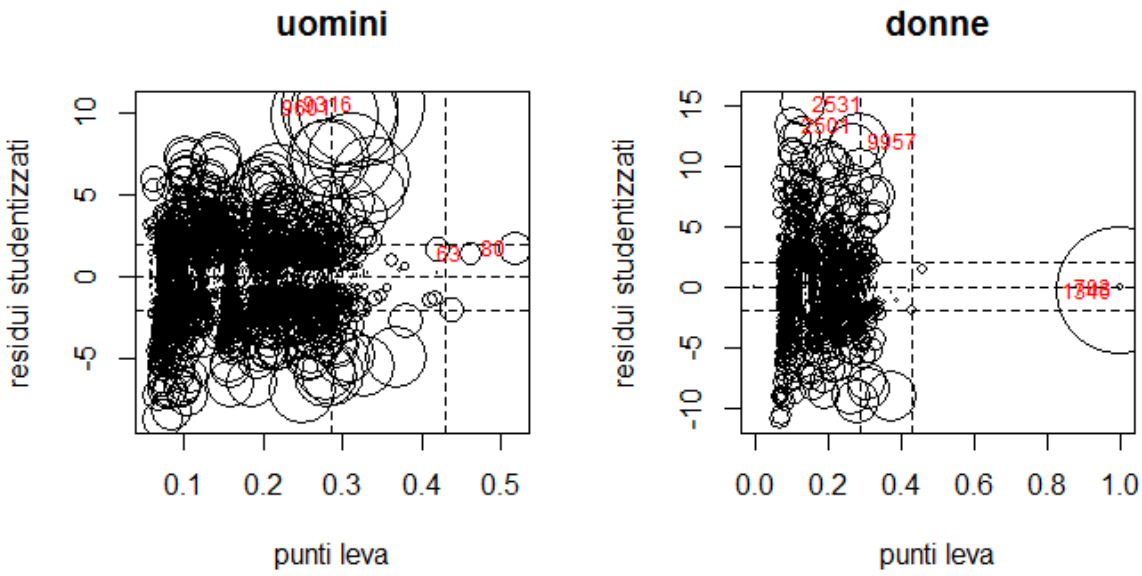

Figura 4.6: Punti influenti per il GLM Poisson

Per la rappresentazione sono stati scelti i due punti con valore maggiore di una delle 3 caratteristiche per semplicità grafica, in realtà sarebbero le bande tratteggiate ad indicarli. I punti rilevati come maggior influenti e le relative misure sono riportati nella tabella 4.2.

Questi punti si riferiscono, in particolare, alle osservazioni che riguardano gli individui con le seguenti caratteristiche:

	Uomini				Donne		
punto	res.stud.	puno leva	dist.Cook	punto	res.stud.	punto leva	dist.Cook
63	1.3820	0.4609	0.0024	703	0.0000	1.0000	0.0014
80	1.7842	0.5177	0.0054	1346	-0.3072	0.9997	0.3927
9316	10.5204	0.3244	0.0777	2501	13.3045	0.1029	0.0307
9601	10.3042	0.2964	0.0647	2531	15.0424	0.1286	0.0503
				9957	11.9369	0.2830	0.0811

Tabella 4.2: Punti influenti del GLM di Poisson

63: uomini con un'età inferiore ad un anno morti nel 2000 per la causa 13, ossia malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico;

80: uomini con un'età compresa tra 1 e 5 anni deceduti nel 2000 per la causa 5, ossia disturbi mentali e comportamentali;

9316: uomini con un'età compresa tra i 65 e i 70 anni morti nel 2013 per la causa 17, ossia altre malattie o causa ignota;

9601: uomini con un'età compresa tra gli 85 e i 90 anni deceduti nel 2013 per la causa 17, ossia altre malattie o causa ignota;

703: donne nel primo anno di vita morte nel 2000 per la causa 5, ossia disturbi mentali e comportamentali;

1346: donne nel primo anno di vita decedute nel 2001 per la causa 5, ossia disturbi mentali e comportamentali;

2501: donne con un'età compresa tra gli 85 e i 90 anni morte nel 2003 per la causa 4, ossia malattie endocrine, nutrizionali e metaboliche;

2531: donne con un'età compresa tra 90 e i 95 anni decedute nel 2003 per la causa 4, ossia malattie endocrine, nutrizionali e metaboliche;

9957: donne con un'età compresa tra gli 85 e i 90 anni morte nel 2013 per la causa 17, ossia altre malattie o malattia ignota.

Il modello nel complesso non è soddisfacente. Si passa ora al GLM con risposta Binomiale Negativa senza procedere ad ulteriori analisi.

4.1.2 Il GLM con risposta Binomiale Negativa

Come già anticipato nel capitolo 3 la distribuzione Binomiale Negativa permette di modellare situazioni in cui la varianza cresce più velocemente della media, cosa che dai grafici dignostici della precedente sezione sembra piuttosto evidente con i dati che si stanno analizzando.

Per implementare questo modello è necessario scaricare la libreria *"MASS"* di R e utilizzare il comando *glm.nb*. La tabella 4.3 riporta i coefficienti degli effetti principali e alcune altre informazioni importanti per poterli interpretare:

	Uomini			Donne		
	Stima	*Std Error*	*p-value*	Stima	*Std Error*	*p-value*
causa 1	-8.8419	0.0808	0.0000	-9.0543	0.0926	0.0000
causa 2	-10.0536	0.1332	0.0000	-10.0125	0.1353	0.0000
causa 3	-10.1771	0.1495	0.0000	-10.4210	0.1791	0.0000
causa 4	-9.1819	0.0855	0.0000	-9.2801	0.0935	0.0000
causa 5	-39.1032	157708.6315	0.9998	-12.7994	1.0015	0.0000
causa 6	-8.5956	0.0667	0.0000	-8.6683	0.0721	0.0000
causa 7	-9.0834	0.0913	0.0000	-9.5259	0.1085	0.0000
causa 8	-10.3396	0.1775	0.0000	-10.2930	0.1955	0.0000
causa 9	-12.0955	0.3776	0.0000	-12.8512	0.5563	0.0000
causa 10	-10.2063	0.1453	0.0000	-10.3979	0.1699	0.0000
causa 11	-10.3815	0.1793	0.0000	-10.8333	0.2187	0.0000
causa 12	-9.4811	0.1137	0.0000	-9.9154	0.1415	0.0000
causa 13	-12.9071	0.6994	0.0000	-12.6083	0.5916	0.0000
causa 14	-11.0895	0.2705	0.0000	-11.8623	0.3840	0.0000
causa 15	-5.5008	0.0271	0.0000	-5.7699	0.0320	0.0000
causa 16	-8.6368	0.0703	0.0000	-8.9239	0.0831	0.0000
causa 17	-8.4036	0.0572	0.0000	-8.5764	0.0654	0.0000
$[1-5)$	-2.1182	0.1335	0.0000	-2.0526	0.1502	0.0000
$[5-10)$	-3.6281	0.2212	0.0000	-3.5031	0.2285	0.0000
$[10-15)$	-3.7878	0.2228	0.0000	-3.6629	0.2509	0.0000
$[15-20)$	-3.1092	0.1693	0.0000	-3.1637	0.1973	0.0000

$[20-25)$	-2.7819	0.1545	0.0000	-3.0509	0.1867	0.0000
$[25-30)$	-1.9411	0.1221	0.0000	-2.4227	0.1551	0.0000
$[30-35)$	-0.8760	0.0999	0.0000	-1.4683	0.1232	0.0000
$[35-40)$	0.0800	0.0910	0.3792	-0.8613	0.1120	0.0000
$[40-45)$	0.2534	0.0892	0.0045	-0.9261	0.1110	0.0000
$[45-50)$	-0.0610	0.0897	0.4967	-1.0276	0.1111	0.0000
$[50-55)$	-0.1031	0.0897	0.2503	-1.0452	0.1107	0.0000
$[55-60)$	0.1160	0.0900	0.1974	-0.6081	0.1086	0.0000
$[60-65)$	0.4231	0.0894	0.0000	-0.1081	0.1058	0.3071
$[65-70)$	0.8814	0.0879	0.0000	0.3689	0.1024	0.0003
$[70-75)$	1.3750	0.0868	0.0000	1.0327	0.0999	0.0000
$[75-80)$	1.9400	0.0860	0.0000	1.6611	0.0986	0.0000
$[80-85)$	2.4821	0.0861	0.0000	2.2297	0.0983	0.0000
$[85-90)$	3.0223	0.0865	0.0000	2.9009	0.0980	0.0000
$[90-95)$	3.4971	0.0877	0.0000	3.4312	0.0981	0.0000
[95 e oltre)	4.0099	0.0953	0.0000	3.9013	0.0996	0.0000
anno	-0.0999	0.0124	0.0000	-0.1031	0.0143	0.0000
anno^2	0.0008	0.0001	0.0000	0.0010	0.0001	0.0000

Tabella 4.3: *Output* parziale del GLM con risposta Binomiale Negativa relativo agli effetti principali diviso per uomini e donne (le celle colorate individuano i *p-value* superiori a 0.01, i cui parametri non sono significativamente diversi da 0)

Da un'attenta analisi si può notare come i parametri stimati tra il modello di Poisson e il Binomiale Negativo siano molto simili. I coefficienti relativi alle cause sono tutti negativi e come per il modello di Poisson il parametro relativo alla causa 5 non è significativamente diverso da 0; nemmeno in questo modello lo è per tutte le interazioni con le altre variabili e gli *Standard Error* a questa causa legati sono ancora una volta troppo alti per non destare sospetti. Anche per quanto riguarda le classi di età non vi sono differenze sostanziali: non sono significative quelle relative agli intervalli [35-40) e [45-60) per gli uomini, mentre per le donne la classe dei [60-65) anni. I parametri relativi agli anni sono ancora una volta negativi per il primo grado e positivi per il secondo.

L'interpretazione dei coefficienti è la stessa del modello di Poisson, infatti la funzione legame è la medesima.

Nel modello Binomiale Negativo vi è la presenza della stima di un ulteriore parametro chiamato "parametro di dispersione", θ, in questo modello tale parametro viene stimato a 465.0 e lo *Standard error* è pari a 20.9 per gli uomini e 328.0 con uno *Standard error* di 13.5 nel modello per le donne.

La devianza del modello Binomiale Negativo è più bassa rispetto al Poisson, in particolare è di 5110 per il modello relativo agli uomini e 4917 per le donne, sempre da confrontare con un χ^2 con 4283 gradi di libertà: situazione migliore rispetto alla precedente.

Per confermare il miglioramento ottenuto con il GLM Binomiale Negativo si passa alle analisi diagnostiche relative. Il primo grafico mostra l'interpolazione tra i residui di devianza e i valori predetti.

Come nel caso precedente diventa difficile analizzare il grafico 4.7 a causa della presenza di valori molto bassi verso l'estremo inferiore dell'ascissa, quindi si passa alla rimozione.

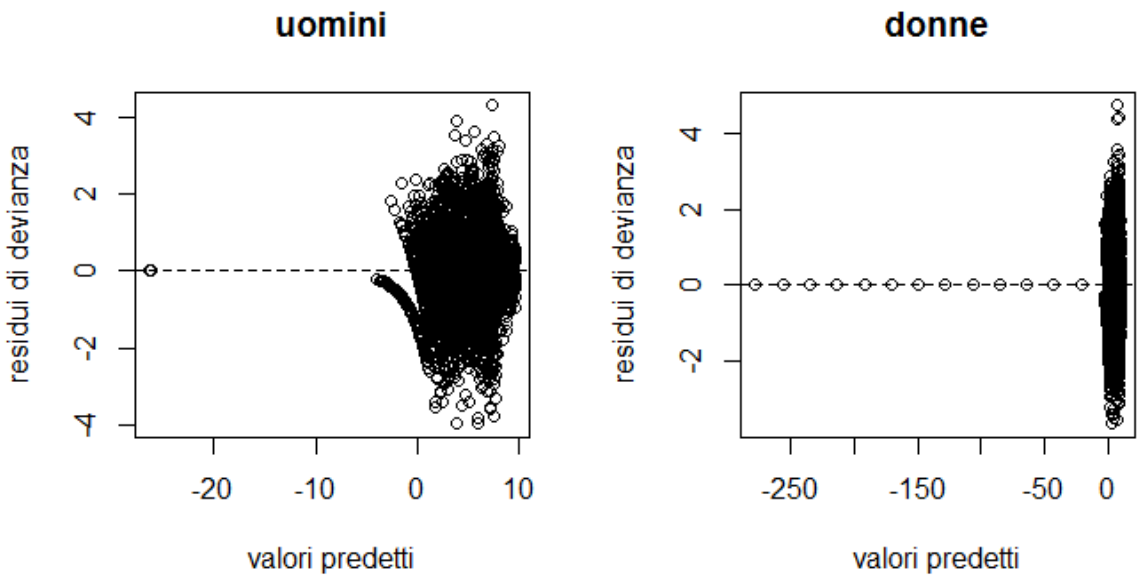

Figura 4.7: Residui di devianza contro i valori predetti per sesso: GLM Binomiale Negativa

Dal grafico 4.8 si può vedere come la situazione sia decisamente migliorata rispetto allo stesso della distribuzione di Poisson; l'utilizzo di un GLM con risposta

Binomiale Negativa permette infatti di cogliere le dinamiche di situazioni in cui sia presente una variabilità più elevata della media. I punti sono più omogenei soprattutto nelle parti destre dei grafici, dove prima si notava la maggiore irregolarità; vi è comunque ancora la presenza di qualche fattore anomalo, infatti sono presenti dei particolari andamenti nell'estremo negativo dei valori predetti (a sinistra dei grafici).

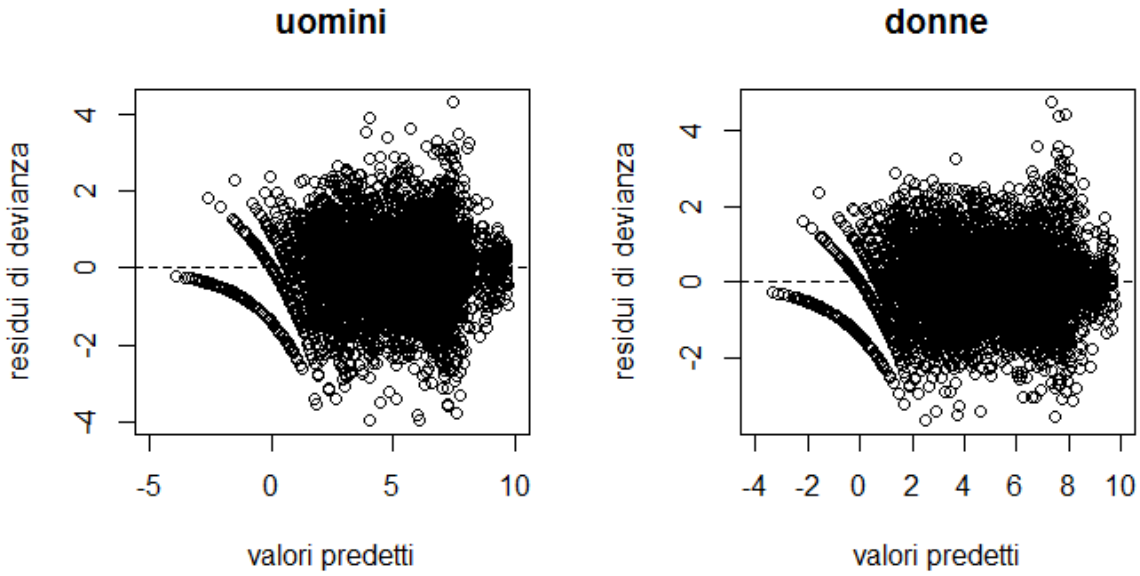

Figura 4.8: Residui di devianza contro valori predetti diviso per sesso (senza la combinazione "causa 5" ed "età [0-1)": GLM Binomiale Negativa

Non sono stati riportati i grafici relativi all'interpolazione tra i valori della proporzione di decessi osservati rispetto a quelli previsti e quello raffigurante l'adeguatezza della funzione legame perchè coincidono con quelli del modello precedente, infatti è stata usata la medesima funzione legame anche per il modello Binomiale Negativo, ossia il logaritmo.

I punti influenti sono invece riportati nel grafico 4.9.

Anche in questo caso, per la rappresentazione, sono stati scelti i due punti con valore maggiore in una delle 3 caratteristiche, questi e le relative misure sono riportati nella tabella 4.4.

Alcuni punti coincidono con quelli individuati dal precedente modello ma hanno in generale misure inferiori rispetto a questo. Anche in questo caso, come nel pre-

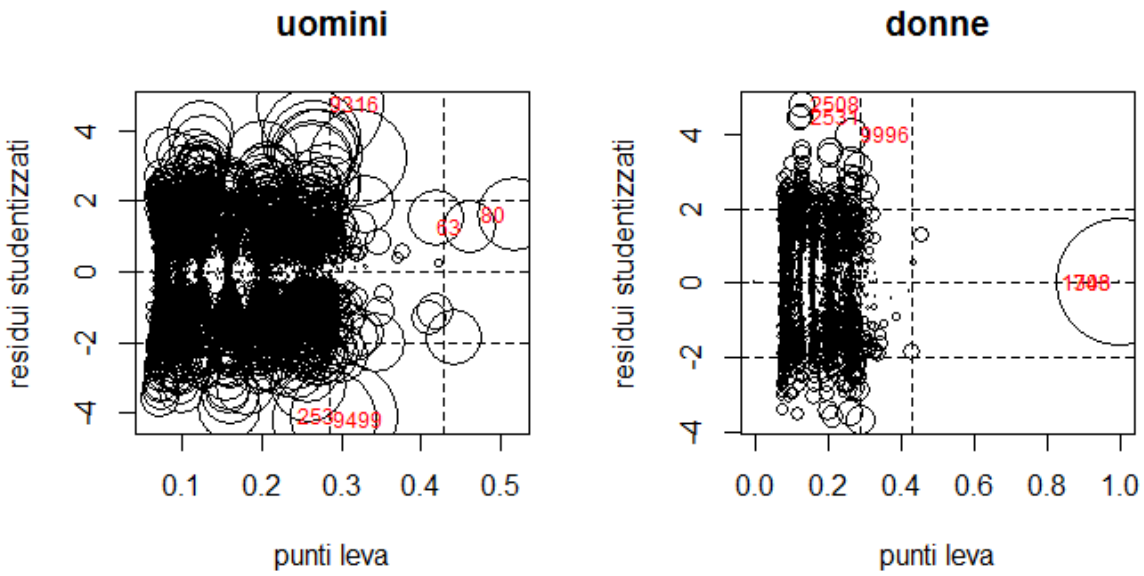

Figura 4.9: Punti influenti per il GLM Binomiale Negativo

	Uomini				Donne		
punto	res.stud.	puno leva	dist.Cook	punto	res.stud.	punto leva	dist.Cook
63	1.2603	0.4610	0.0026	703	0.0000	1.0000	0.1375
80	1.6279	0.5174	0.0054	1346	-0.0019	0.9999	0.0000
253	-4.1128	0.3026	0.0111	2508	4.7904	0.1257	0.0061
9316	4.7328	0.2742	0.0154	2531	4.4847	0.157	0.0053
9499	-4.2199	0.2779	0.0103	9996	3.9810	0.2616	0.0099

Tabella 4.4: Punti influenti del GLM Binoamiale Negativo

cedente, si può notare che nel grafico relativo alle donne vi sia la presenza di punti influenti che hanno residui studentizzati nella norma ma elevati valori di punti leva (in questo casi i punti 703 e 1346).

Questi punti si riferiscono, in particolare, alle osservazioni che riguardano gli individui con le seguenti caratteristiche:

63: uomini con un'età inferiore ad un anno morti nel 2000 per la causa 13, ossia malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico;

80: uomini con un'età compresa tra 1 e 5 anni deceduti nel 2000 per la causa 5, ossia disturbi mentali e comportamentali;

253: uomini con un'età compresa tra i 40 e i 45 anni morti nel 2000 per la causa 1, ossia malattie infettive;

9316: uomini con un'età compresa tra i 65 e i 70 anni deceduti nel 2013 per la causa 17, ossia altre malattie o causa ignota;

9499: uomini con un'età superiore ai 95 anni morti nel 2013 per la causa 6, ossia malattie del sistema nervoso e degli organi di senso;

703: donne nel primo anno di vita decedute nel 2000 per la causa 5, ossia disturbi mentali e comportamentali;

1346: donne nel primo anno di vita morte nel 2001 per la causa 5, ossia disturbi mentali e comportamentali;

2508: donne con un'età superiore ai 95 anni morte nel 2003 per la causa 4, ossia malattie endocrine, nutrizionali e metaboliche;

2531: donne con un'età compresa tra 90 e i 95 anni decedute nel 2003 per la causa 4, ossia malattie endocrine, nutrizionali e metaboliche;

9996: donne con un'età compresa tra gli 70 e i 75 anni morte nel 2013 per la causa 17, ossia altre malattie o casua ignota.

Il modello GLM con risposta Binomiale Negativa è soddisfacente, sembra cogliere bene la variabilità dei dati ed ha una Devianza Residua piuttosto contenuta. Si passa ora all'analisi dei dati attraverso i modelli *Zero Inflated*, descritti anch'essi nel capitolo 3.

4.2 I modelli Zero Inflated

I modelli ad inflazione di zeri sono caratterizzati dalla presenza di una variabile risposta che si suppone essere la mistura di due distribuzioni, una legata al verificarsi dell'esito positivo e l'altra, dato che si è verificato l'evento, al valore che questo assume.

A differenza dei modelli GLM sono stati considerati, nei modelli *Zero Inflated*, per quanto riguarda la parte relativa ai valori positivi (ossia alla distribuzione di Poisson nello ZIP e alla Binomiale Negativa nello ZINB), solo gli effetti principali e le interazioni di primo grado, tra le cause e le classi di età: questo perchè le altre interazioni non sono risultate utili per l'adattamento del modello ai dati. Per quanto attiene alla parte relativa al verificarsi dell'evento (quindi alla distribuzione Binomiale) è stata inserita la sola intercetta. L'interpretazione dei coefficienti è la stessa data nei modelli precedenti, a patto di considerare, ove significativa, la sezione relativa all'inflazione di zeri. Anche in questo caso, per la parte riguardante il valore assunto dalla variabile (per le distribuzioni di Poisson e Binomiale Negativa), una volta che l'evento si è verificato, non è stata considerata l'intercetta, in modo da avere un coefficiente per ogni causa.

4.2.1 Il modello ZIP

Per implementare questo tipo di modello è necessario ricorrere alla libreria *pscl* e utilizzare il comando *zeroinfl{....}*.

I parametri stimati relativi agli effetti principali del modello *Poisson Zero Inflated* sono riportati nella tabella 4.5. Rispetto al GLM di Poisson si può notare come vi sia la presenza di un ulteriore parametro, in fondo alla tabella, separato dai precedenti da una linea: si tratta della parte del modello relativa all'inflazione di zeri. Anche in questo caso vi è la presenza di due modelli separati, uno per i dati relativi agli uomini e uno per le donne.

	Uomini			Donne		
	Stima	*Std Error*	*p-value*	Stima	*Std Error*	*p-value*
causa 1	3.4559	0.0468	0.0000	3.1757	0.0535	0.0000

causa 2	2.5730	0.0727	0.0000	2.5301	0.0739	0.0000
causa 3	2.1988	0.0877	0.0000	1.7514	0.1091	0.0000
causa 4	3.5832	0.0439	0.0000	3.3936	0.0480	0.0000
causa 5	-11.3322	76.0771	0.8816	-2.6794	1.0000	0.0074
causa 6	4.0917	0.0340	0.0000	3.9565	0.0362	0.0000
causa 7	3.1660	0.0541	0.0000	2.9731	0.0592	0.0000
causa 8	1.6491	0.1155	0.0000	1.2909	0.1374	0.0000
causa 9	0.3758	0.2182	0.0851	-0.3768	0.3162	0.2334
causa 10	2.3617	0.0808	0.0000	1.9553	0.0985	0.0000
causa 11	1.6518	0.1165	0.0000	1.3459	0.1336	0.0000
causa 12	2.7055	0.0708	0.0000	2.2184	0.0864	0.0000
causa 13	-1.2822	0.5000	0.0103	-0.8876	0.4082	0.0297
causa 14	0.6989	0.1857	0.0002	0.0932	0.2500	0.7094
causa 15	7.2015	0.0072	0.0000	6.9105	0.0083	0.0000
causa 16	3.8781	0.0379	0.0000	3.5432	0.0446	0.0000
causa 17	4.6741	0.0255	0.0000	4.3733	0.0294	0.0000
$[1-5)$	-0.4042	0.0739	0.0000	-0.3457	0.0832	0.0000
$[5-10)$	-1.8346	0.1261	0.0000	-1.4483	0.1227	0.0000
$[10-15)$	-1.7683	0.1225	0.0000	-1.8477	0.1451	0.0000
$[15-20)$	-1.0940	0.0934	0.0000	-1.1546	0.1093	0.0000
$[20-25)$	-0.8777	0.0863	0.0000	-0.9351	0.1008	0.0000
$[25-30)$	-0.2699	0.0711	0.0001	-0.5568	0.0887	0.0000
$[30-35)$	0.6125	0.0581	0.0000	0.2696	0.0711	0.0001
$[35-40)$	1.4997	0.0517	0.0000	0.8748	0.0637	0.0000
$[40-45)$	1.9349	0.0500	0.0000	1.0900	0.0619	0.0000
$[45-50)$	2.0324	0.0497	0.0000	1.2241	0.0609	0.0000
$[50-55)$	2.0759	0.0496	0.0000	1.3265	0.0602	0.0000
$[55-60)$	2.1387	0.0495	0.0000	1.6014	0.0587	0.0000
$[60-65)$	2.2563	0.0492	0.0000	1.7749	0.0579	0.0000
$[65-70)$	2.4062	0.0488	0.0000	2.1355	0.0566	0.0000
$[70-75)$	2.7457	0.0483	0.0000	2.7036	0.0553	0.0000

$[75-80)$	3.1578	0.0478	0.0000	3.2905	0.0545	0.0000
$[80-85)$	3.3454	0.0476	0.0000	3.6933	0.0542	0.0000
$[85-90)$	3.2001	0.0477	0.0000	3.9010	0.0541	0.0000
$[90-95)$	2.6369	0.0484	0.0000	3.6785	0.0542	0.0000
[95 e oltre)	1.5743	0.0514	0.0000	3.1187	0.0547	0.0000
anno	0.0045	0.0001	0.0000	0.0062	0.0001	0.0000
intercetta	-22.2274	0.9777	0.000	-33.98	306.19	0.912

Tabella 4.5: *Output* parziale dello ZIP relativo agli effetti principali diviso per uomini e donne (le celle colorate individuano i *p-value* superiori a 0.01, i cui parametri non sono quindi significativamente diversi da 0). La prima parte (prima della riga) è riferita alla distribuzione di Poisson mentre la seconda alla Binomiale.

Analizzando nel dettaglio la tabella contenente gli effetti principali dei due modelli ZIP applicati ai dati, si può notare come vi siano delle differenze piuttosto importanti rispetto al semplice modello GLM di Poisson. Si rivela ora necessario un discorso separato per uomini e donne, infatti per i primi l'utilizzo di un modello *Zero Inflated* risulta adeguato, il parametro relativo è infatti significativamente diverso da 0, cosa che invece non accade per le donne. Per i primi bisogna focalizzare l'attenzione sui parametri relativi alle cause, i quali potrebbero trarre in inganno, in quanto non sono più tutti negativi come in precedenza, quindi si potrebbe dedurre che queste abbiano un impatto positivo sul logaritmo della proporzione di decessi; in realta bisogna considerare anche l'intercetta relativa alla parte del modello ad inflazione di zeri, considerando la quale si ottengono dei coefficienti in linea con i precedenti. Di maggior rilevanza è invece la presenza di più cause con parametro non significativamente diverso da 0 rispetto al modello precedente, in particolare oltre alla 5, attinente ai decessi per disturbi mentali e comportamentali, (che presenta ancora *Standard error* troppo elevati) vi è anche la 9 (disturbi dell'apparato circolatorio), discorso a parte andrebbe fatto per la 13 (malattie della pelle, del tessuto sottocutaneo, connettivo ed del sistema muscolo-scheletrico), che mostra un *p-value* attorno alla soglia critica scelta. Queste ultime due cause hanno comunque parametri significativi quando vengono considerate nell'interazione con alcune classi di età, a differenza della 5 che risulta sempre non significativa. Per quanto riguarda gli effetti

principali delle classi di età e dell'anno considerato questi sono tutti significativi. Per il modello relativo alle donne si può comprendere, dalla non significatività del coefficiente relativo al modello *Zero Inflated*, che non è utile utilizzare un modello ZIP. Curiosa è comunque la non significatività dei parametri legati alle cause 9, 13 e 14, mentre per quanto riguarda le classi di età i coefficienti sono tutti diversi da 0. A causa di questa non significatività del parametro relativo all'inflazione di zeri per le donne, i grafici diagnostici per queso gruppo saranno presenti ma poco rilevanti.

Il primo grafico, il 4.10, quello relativo all'interpolazione tra i valori stimati dal modello e le proporzioni di decessi veramente osservate, non è per nulla rassicurante, i punti dovrebbero giacere sulla bisettrice, invece sia per gli uomini che per le donne questi sono piuttosto lontani dalla situazione ideale.

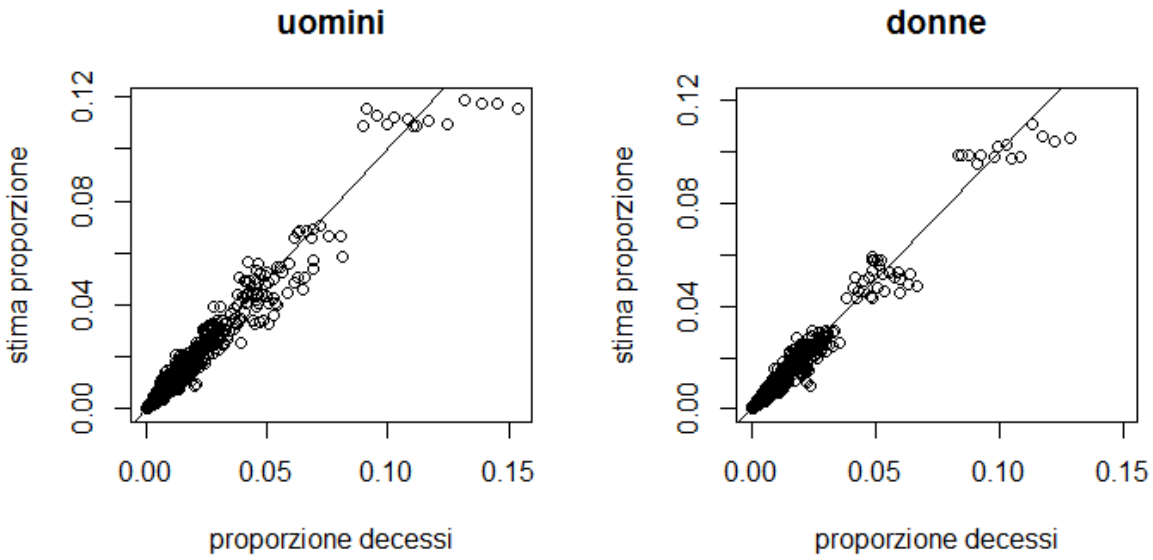

Figura 4.10: Stima della proporzione contro valore osservato, diviso per sesso: ZIP

Anche il grafico 4.11 relativo ai residui di Pearson contro i valori predetti non è soddisfacente; i punti dovrebbero essere disposti in modo piuttosto omogeneo, invece all'aumentare dei valori predetti il *range* dei residui tende a diminuire. Un grafico di questo tipo fa pensare che la variabilità dei dati non sia colta come dovrebbe, sembra che questa cresca meno velocemente della media.

Dalle analisi diagnostiche e dalla non significatività del coefficiente associato all'inflazione di zeri per le donne, si intuisce che i modelli *Zero Inflated* non siano

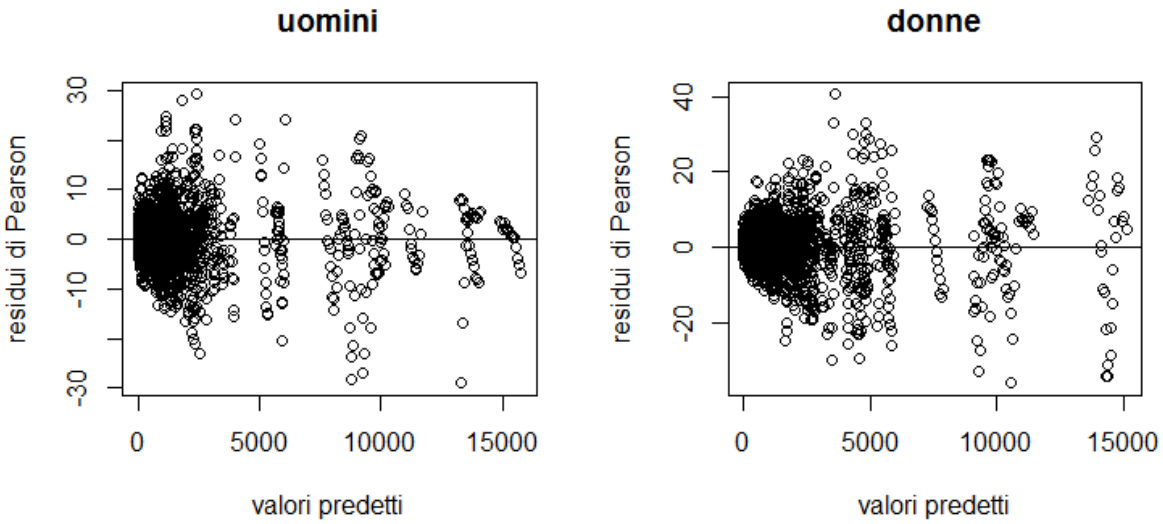

Figura 4.11: Residui di Pearson contro i valori predetti diviso per sesso: ZIP

adeguati alla modellazione dei dati considerati, questo potrebbe essere causato dalla presenza piuttosto ridotta di veri e propri 0 all'interno dei dati.

4.2.2 Il modello ZINB

Si passa ora allo *Negative Binomial Zero Inflated*, per il quale è necessario anche in questo caso ricorrere alla libreria *pscl* e utilizzare il comando *zeroinfl{...., dist="negbin"}*. La tabella contenente i coefficienti stimati relativi agli effetti principali è la 4.6.

	Uomini			Donne		
	Stima	*Std Error*	*p-value*	Stima	*Std Error*	*p-value*
causa 1	3.5115	0.0651	0.0000	3.2548	0.0697	0.0000
causa 2	2.6308	0.0857	0.0000	2.6120	0.0864	0.0000
causa 3	2.2562	0.0987	0.0000	1.8328	0.1179	0.0000
causa 4	3.6417	0.0630	0.0000	3.4761	0.0656	0.0000
causa 5	-11.2785	76.2590	0.8824	-2.5939	0.9995	0.0095
causa 6	4.1494	0.0566	0.0000	4.0384	0.0575	0.0000
causa 7	3.2218	0.0705	0.0000	3.0544	0.0742	0.0000
causa 8	1.7061	0.1240	0.0000	1.3711	0.1444	0.0000

causa 9	0.4341	0.2229	0.0514	-0.2942	0.3194	0.3569
causa 10	2.4195	0.0926	0.0000	2.0366	0.1082	0.0000
causa 11	1.7070	0.1244	0.0000	1.4278	0.1409	0.0000
causa 12	2.7540	0.0867	0.0000	2.2991	0.0972	0.0000
causa 13	-1.2250	0.5023	0.0147	-0.8052	0.4107	0.0499
causa 14	0.7565	0.1911	0.0001	0.1753	0.2540	0.4899
causa 15	7.2591	0.0458	0.0000	6.9929	0.0454	0.0000
causa 16	3.9347	0.0590	0.0000	3.6237	0.0631	0.0000
causa 17	4.7350	0.0520	0.0000	4.4588	0.0535	0.0000
$[1-5)$	-0.4022	0.0976	0.0000	-0.3429	0.1042	0.0010
$[5-10)$	-1.8319	0.1412	0.0000	-1.4448	0.1378	0.0000
$[10-15)$	-1.7656	0.1381	0.0000	-1.8449	0.1581	0.0000
$[15-20)$	-1.0918	0.1130	0.0000	-1.1515	0.1261	0.0000
$[20-25)$	-0.8754	0.1073	0.0000	-0.9316	0.1188	0.0000
$[25-30)$	-0.2690	0.0954	0.0048	-0.5547	0.1087	0.0000
$[30-35)$	0.6092	0.0862	0.0000	0.2686	0.0948	0.0046
$[35-40)$	1.4945	0.0820	0.0000	0.8721	0.0894	0.0000
$[40-45)$	1.9327	0.0810	0.0000	1.0907	0.0881	0.0000
$[45-50)$	2.0346	0.0808	0.0000	1.2278	0.0875	0.0000
$[50-55)$	2.0789	0.0807	0.0000	1.3318	0.0870	0.0000
$[55-60)$	2.1427	0.0806	0.0000	1.6078	0.0859	0.0000
$[60-65)$	2.2614	0.0804	0.0000	1.7807	0.0854	0.0000
$[65-70)$	2.4084	0.0802	0.0000	2.1386	0.0845	0.0000
$[70-75)$	2.7467	0.0799	0.0000	2.7044	0.0836	0.0000
$[75-80)$	3.1602	0.0796	0.0000	3.2925	0.0831	0.0000
$[80-85)$	3.3509	0.0795	0.0000	3.7003	0.0829	0.0000
$[85-90)$	3.2073	0.0796	0.0000	3.9099	0.0828	0.0000
$[90-95)$	2.6427	0.0800	0.0000	3.6842	0.0829	0.0000
[95 e oltre)	1.5797	0.0818	0.0000	3.1267	0.0833	0.0000
anno	-0.0044	0.0007	0.0000	-0.0066	0.0008	0.0000
$log(\theta)$	3.5630	0.0273	0.0000	3.5911	0.0283	0.0000

intercetta	-22.476	1.057	0.0000	-30.70	58.32	0.599

Tabella 4.6: *Output* parziale del modello ZINB relativo agli effetti principali diviso per uomini e donne (le celle colorate individuano i *p-value* superiori a 0.01, i cui parametri non sono quindi significativamente diversi da 0). La prima parte (prima della riga) è riferita alla distribuzione Binomiale Negativa mentre la seconda alla Binomiale.

Come per il modello ZIP anche in questo caso il modello ad inflazione di zeri non sembra adeguato ai dati riguardanti le donne, mentre in quello degli uomini, il parametro relativo alla parte ad inflazione di zeri è significativo. In quest'ultimo le cause con coefficiente non significativamente diverso da 0 sono ancora una volta la 5 (che anche in questo modello presenta *standard error* molto elvati) e la 9; la 13 mostra nuovamente un *p-value* vicino alla soglia. Ad eccezione della prima, le cause 9 e 13 sono significative quando sono considerate nelle interazioni con le classi di età.

Trattandosi di una distribuzione Binomiale Negativa anche in questo caso è presente il parametro θ, con trasformata pari alla funzione legame, che viene interpretato come il parametro di dispersione della distribuzione, riportato nella scala originale attraverso la trasformata inversa ossia l'esponenziale, si ottiene 35.26885 per gli uomini e 36.27396 per le donne: questi valori indicano che una distribuzione Binomiale Negativa è più adatta della distribuzione di Poisson per cogliere la sovra-dispersione presente.

Il grafico diagnostico, relativo all'interpolazione tra i residui di Pearson e i valori predetti, il 4.12, non è per nulla confortante, mostra come la situazione sia anche peggiore rispetto allo ZIP: nell'estremo inferiore dei valori predetti i punti sono piuttosto schiacciati, il range dei residui è abbastanza allargato e via via che questi valori in ascissa crescono il range dell'ordinata diminuisce. In effetti la distribuzione Binomiale Negativa permette di modellare situazioni in cui la varianza cresce più velocemente della media, e non più lentamente, come si era già intuito dal grafico 4.11 relativo al modello ZIP.

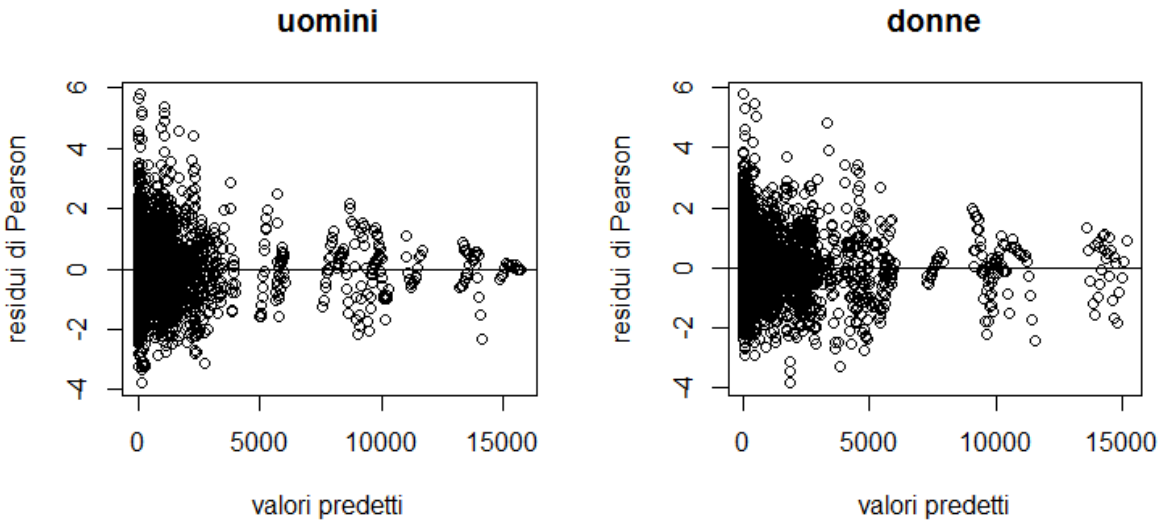

Figura 4.12: Residui di Pearson contro i valori predetti diviso per sesso: ZINB

4.3 Confronto tra i modelli e scelta del modello finale

Il confronto tra i modelli stimati circa l'adeguatezza degli stessi ai dati viene fatto sulla base delle analisi diagnostiche condotte e sul valore dell'AIC di ogni modello. Per quanto riguarda il primo aspetto i grafici che presentano meno problemi sono quelli relativi al secondo modello stimato, ossia al GLM con risposta Binomiale Negativa.

Relativamente all'AIC, ossia il criterio di informazione di Akaike, questo rappresenta una valutazione della perdita attesa di efficacia predittiva [Pace e Salvan 2001] ed è formulato come:

$$AIC(F_p) = 2 \cdot p - 2 \cdot l(\hat{\theta}^p, y)$$

dove:

- p è il numero di parametri stimati nel modello;
- $l(\hat{\theta}^p, y)$ è la log-verosimiglianza del modello.

Questo criterio seleziona, come modello migliore, quello con valore minore, in particolare nella tabella 4.7 sono riportati i valori dei modelli fin qui descritti.

Il modello che presenta un valore minore è quello relativo al GLM con risposta Binomiale Negativa, anche se non è di molto inferiore rispetto al GLM di Poisson

	Uomini	Donne
modello	AIC	AIC
GLM Poisson	45155	47862
GLM Binomiale Negativo	41965	40858
ZIP	140568	148435
ZINB	47592	45465

Tabella 4.7: Criterio di Akaike per i modelli stimati, le celle colorate indicano i valori più bassi

e allo ZINB. Bisogna però considerare che mentre nei modelli GLM sono presenti le interazioni fino al secondo ordine tra la variabile relativa alla causa, quella della classe di età e l'anno di rilevazione, nei modelli *Zero Inflated* sono state inserite solo le interazioni di primo grado tra queste prime due variabili. Fatta questa osservazione pare più opportuno scegliere come modello migliore quello relativo allo ZINB per gli uomini e il GLM con risposta Binomiale Negativa per le donne. Per queste ultime infatti il coefficiente relativo alla parte *Zero Inflated* non è risultato significativamente diverso da 0. In appendice è riportata la tabella, la numero A.2, che contiene i parametri relativi alle interazioni di primo e secondo grado, oltre agli effetti principali, per il modello scelto per spiegare al meglio il fenomeno oggetto di studio relativo alle donne.

Per gli uomini però si reputa comunque necessaria una modifica rispetto al modello riportato nella sezione 4.2.2, infatti, come già discusso, la causa 5, ossia i disturbi mentali e comportamentali, ha uno *Standard error*, per tutti i gradi di interazione, troppo elevato per essere ignorato. Per questo motivo è stato nuovamente stimato il modelo ZINB senza tale causa e gli effetti principali sono riportati in tabella 4.8.

	Stima	*Std.Error*	*p-value*
causa 1	3.5154	0.0650	0.0000
causa 2	2.6351	0.0857	0.0000
causa 3	2.2607	0.0987	0.0000
causa 4	3.6458	0.0630	0.0000

causa 6	4.1537	0.0566	0.0000
causa 7	3.2255	0.0705	0.0000
causa 8	1.7097	0.1240	0.0000
causa 9	0.4381	0.2229	0.0493
causa 10	2.4237	0.0926	0.0000
causa 11	1.7117	0.1244	0.0000
causa 12	2.7581	0.0866	0.0000
causa 13	-1.2202	0.5021	0.0151
causa 14	0.7603	0.1912	0.0001
causa 15	7.2633	0.0458	0.0000
causa 16	3.9386	0.0590	0.0000
causa 17	4.7397	0.0520	0.0000
$[1-5)$	-0.4019	0.0975	0.0000
$[5-10)$	-1.8313	0.1412	0.0000
$[10-15)$	-1.7649	0.1380	0.0000
$[15-20)$	-1.0914	0.1130	0.0000
$[20-25)$	-0.8750	0.1072	0.0000
$[25-30)$	-0.2690	0.0954	0.0048
$[30-35)$	0.6089	0.0861	0.0000
$[35-40)$	1.4938	0.0820	0.0000
$[40-45)$	1.9322	0.0809	0.0000
$[45-50)$	2.0349	0.0808	0.0000
$[50-55)$	2.0793	0.0807	0.0000
$[55-60)$	2.1432	0.0806	0.0000
$[60-65)$	2.2621	0.0804	0.0000
$[65-70)$	2.4086	0.0802	0.0000
$[70-75)$	2.7469	0.0798	0.0000
$[75-80)$	3.1605	0.0796	0.0000
$[80-85)$	3.3517	0.0795	0.0000
$[85-90)$	3.2085	0.0795	0.0000
$[90-95)$	2.6435	0.0800	0.0000

[95 e oltre)	1.5804	0.0818	0.0000
anno	-0.0051	0.0008	0.0000
$log(\theta)$	3.5641	0.0282	0.0000
intercetta	-22.4130	1.0548	0.0000

Tabella 4.8: *Output* parziale del modello ZINB relativo agli effetti principali per gli uomini senza la causa 5 (le celle colorate individuano i *p-value* superiori a 0.01 i cui parametri non sono quindi significativamente diversi da 0). La prima parte (prima della riga) è riferita alla distribuzione Binomiale Negativa mentra la seconda alla Binomiale

La tabella con i parametri e gli *Standard error* degli effetti principali e di tutte le interazioni è la numero A.1 in appendice.

Come accadeva nel modello ZINB con anche la causa 5, il coefficiente relativo alla causa 9, relativa ai decessi per altri disturbi dell'apparato circolatorio, risulta non significativamente diversa da 0 mentre quello della causa 13, che considera i morti per malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico, è appena superiore alla soglia. Anche tutti gli altri parametri e i grafici diagnostici restano pressoché uguali a quelli del modello con la causa 5. Anche il parametro di dispersione, θ ha un valore praticamente identico al precedente.

Cambiamento più rilevante si può invece notare nel valore dell'AIC, questo infatti diminuisce rispetto a quello relativo al precedente modello: se non si considera la causa 5, nell'implementazione dello ZINB, il valore raggiunto da tale criterio scende infatti a 44911 (rispetto al 47592, riportato nella tabella 4.7 relativo allo stesso modello con anche la causa 5), questo è però dovuto alla differenza in termini di numerosità dei dati considerati.

Il modello scelto che meglio si adatta ai dati relativi agli uomini per spiegare il fenomeno oggetto dello studio è quindi in modello ZINB a cui viene però rimossa la causa 5, evidentemente per cogliere al meglio le dinamiche che caratterizzano i decessi per questa causa sarebbe necessario un diverso modello statistico.

4.3.1 Interpretazione

Come ampiamente discusso le dinamiche che intercorrono, relative alla mortalità per causa, distinguendo tra uomini e donne sono differenti e piuttosto complesse. Lo stesso tipo di modello scelto, tra quelli proposti, che si suppone spieghi meglio queste dinamiche è differente nei due gruppi di dati considerati. Detto ciò, si vedranno in questa sezione ulteriori differenze e similarità che contraddistinguono questi due casi.

La differenza più marcata è senza dubbio quella relativa al modello scelto, per le donne un modello ad inflazione di zeri risulta poco adeguato a spiegare il fenomeno, mentre per gli uomini questo risulta una scelta consona. Il modello relativo agli uomini è più parsimonioso di quello per le donne, non sono infatti state considerate le interazioni di secondo grado nel primo, mentre nel secondo sì. Differenza non da meno è data dalla presenza della causa di morte relativa a disturbi mentali e comportamentali: per gli uomini crea problema con tutti i modelli sviluppati, infatti si è deciso di toglierla, mentre per le donne questa non provoca particolari situazioni preoccupanti.

Diversità importanti sono presenti anche negli effetti principali delle covariate e quindi di conseguenza anche sulle stime delle proporzioni di decessi per determinate caratteristiche degli individui. Per quanto riguarda la significatività relativa alle interazioni di primo grado[2] (le cui tabelle con i rispettivi parametri stimati sono in appendice), per comodità non è stato riportato in ogni confronto la categoria di riferimento che è sempre per la classe di età quella realtiva ai neonati che non hanno ancora compiuto un anno di età mentre per le cause di morte la numero 1 ossia le malattie infettive:

- la causa 2, ossia neoplasma, ha coefficienti diversi da 0 sia per le donne che per gli uomini per tutte le classi di età, ci sono quindi differenze significative nei confronti della classe di riferimento;
- la causa 3, relativa alle malattie del sangue non ha un comportamento diverso se si considerano le donne, in particolare dalla classe dei [15-20) anni in su,

[2]I commenti relativi agli effetti principali sono già stati fatti e sono presenti nelle sezioni precedenti di questo capitolo, dove i modelli sono stati presentati.

mentre per gli uomini questi coefficienti sono tutti diversi da 0 eccetto per le classi [20-30) e dai 90 anni in su;

- per la causa 4, malattie endocrine, nutrizionali o metaboliche, non ci sono grosse differenze tra uomini e donne, tendono in entrambi i casi a non essere significative le prime classi di età (da 1 a 5 anni per gli uomini, e fino ai 10 per le donne) e per età centrali (dai 15 ai 30 e dai 50 ai 60 anni per gli uomini e dai 25 ai 30 e dai 45 ai 50 anni per le donne);
- per la causa 6, malattie del sistema nervoso e degli organi di senso, gli uomini presentano coefficienti non significativi in età più alte rispetto alle donne (i primi per la macro classe dai 60 agli 80 anni, mentre le seconde dai 40 ai 60 anni), oltre alle classi estreme, sia verso quelle più basse che quelle più alte;
- la causa 7, malattie cardiache, mostra coefficienti significativi in tutte le classi sia per uomini che per donne eccetto fino ai 10 anni;
- per la causa 8, malattie cerebrovascolari, si notano invece due situazioni differenti, per gli uomini l'unica classe di età con coefficiente non significativamente diverso da 0 è quella relativa all'intervallo [1-5) anni mentre per le donne abbiamo le classi dai 5 ai 20 anni e la classe [35-40);
- la causa 9, relativa ad altri disturbi dell'apparato circolatorio mostra una situazione identica alla 7;
- anche per la causa 10, malattie respiratorie acute, la significatività tra i coefficienti dei due modelli è molto simile, sono diversi da 0 quelli relativi alle classi dai 55 anni in poi per gli uomini e dai 65 in poi per le donne;
- per la causa 11, altre malattie respiratorie non acute i coefficienti non significativi sono molto simili tra uomini e donne (per i primi per le classi [1-5) e [35-40) anni mentre per le donne [1-10) e [30-40) anni);
- anche per la causa 12, relativa alle malattie del sistema digerente, non ci sono grosse differenze in termini di non signifiatività tra uomini e donne, per i primi non lo sono la macro classe [1-20) anni mentre per le seconde le classi [1-10) e [15-35);
- per la causa 13, invece, relativa ai morti per malattie della pelle, del tessuto sottocutaneo, connettivo e del sistema muscolo-scheletrico, vi è una notevole

differenza, per gli uomini tutti i coefficienti sono diversi da 0 mentre per le donne non lo solo quelli appartenenti alle classi [1-10) e [30-45);

- per la causa 14, decessi per malattie del sistema genitourinario e complicanze legate alla gravidanza, vi sono importanti differenze, per gli uomini i coefficienti non significativi sono quelli relativi alle classi [1-25) e [30-45) anni, mentre per le donne la macro classe [1-15);
- sia per la causa 15, relativa ai decessi per condizioni originarie del periodo perinatale e anomalie o malformazioni congenite, che per la 16, ossia cause esterne, tutti i coefficienti sia per gli uomini che per le donne sono diversi da 0. La motivazione sottostante è da ricercarsi nella natura della causa stessa, in particolare per la prima questa, come già visto colpisce proprio la categoria di riferimento, i neonati, quindi per tutte le altre classi di età l'influenza di queste nella proporzione di decessi sarà rilevante, discorso opposto per la seconda causa qui analizzata;
- per la causa 17, altre malattie o causa ignota, le situazioni tra uomini e donne non sono molto diverse, per i primi le classi con coefficienti non significativi sono [5-15), [35-75) e [85-95), mentre per le donne [1-15), [30-45) e [55-80).

Dopo questa rassegna sulla significatività dei coefficienti stimati dai due diversi modelli si vedranno che impatto hanno le varie covariate sulla stima della proporzione di decessi, per determinate caratteristiche; per fare ciò sono stati creati dei grafici, sempre distintamente per uomini e donne[3]. Sono stati riportati gli anni 2000, 2006 e 2013 come rappresentativi dei 13 anni analizzati.

Il primo grafico, il 4.13, si riferisce all'anno 2000, e mostra come le proporzioni stimate per gli uomini, rispetto a quelle delle donne, siano generalmente più alte, soprattutto dai 50-60 anni in poi; il valore più elevato si osserva per la causa 7, che raggruppa i decessi per malattie cardiache, dove per gli uomini raggiunge circa il 16%, nella classe delle persone dai 95 anni in su, questa raggiunge quasi sempre le proporzioni più alte anche nelle altre classi di età. Anche per le donne è la classe con valori più alti, ma che raggiungono il massimo intorno all'11%, nell'età più

[3]Per maggiori dettagli sui coefficienti dei vari modelli è possibile fare riferimento alle due tabelle A.1 e A.2 in appendice.

alta considerata. La seconda causa per valore, negli uomini è la 2, ossia neoplasma, mentre nelle donne, questa, ha un valore contento anche per l'ultima classe di età considerata; la seconda causa per valore più alto per le donne della classe 95 anni e oltre è rappresentata dalla 17, ossia altre o causa ignota, che raggiunge circa il 5%, mentre per gli uomini è intorno al 7%.

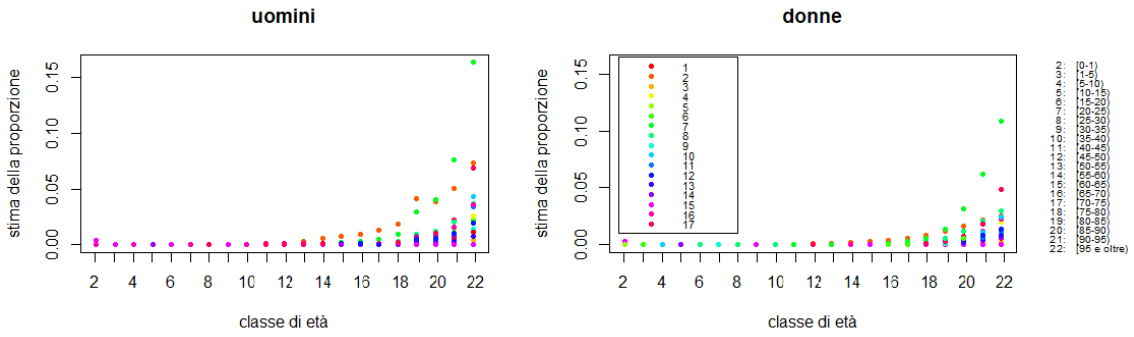

Figura 4.13: Stima della proporzione di decessi per classe di età, causa e sesso: anno 2000

Per meglio comprendere la situazione per valori bassi delle proporzioni stimate, nel grafico 4.14, è riportata un'espansione dello stesso grafico per questi valori.

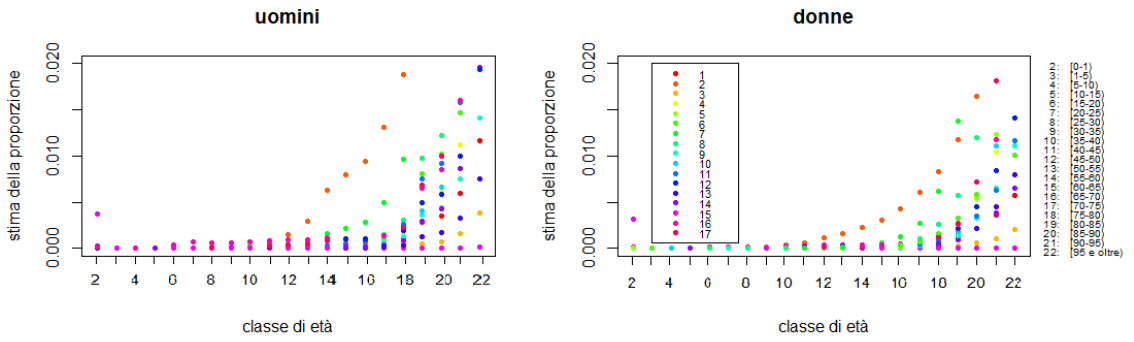

Figura 4.14: Approfondimento per valori bassi della stima della proporzione di decessi per classe di età, causa e sesso: anno 2000

Si può vedere che, sia per gli uomini che per le donne, la causa 2, neoplasma, ha una proporzione di decessi importante (nel grafico precedente per le donne non era chiara questa situazione), ma mentre per le donne questa ha una crescita più lenta, all'aumentare della classe di età, per gli uomini la crescita è più drastica a partire dalla classe [50-55). La causa 15, come era possibile aspettarsi, ha valori alti solo per la classe [0-1) anni, si tratta infatti di condizioni originarie del periodo perinatale e

anomalie o malformazioni congenite, quindi malattie legate ai neonati. Mentre per quasi tutte le altre cause i valori stimati sono prossimi allo 0 fino a età medio basse, quali 20 anni, per gli uomini e centrali, quali 45 anni, per le donne.

Le cause che invece hanno un valore superiore al 2%, oltre a quelle già discusse e visibili chiaramente dal grafico complessivo per il 2000, sono per gli uomini, e sempre per classi di età elevata, in ordine decrescente la 10, malattie respiratorie acute, la 8, malattie cerebrovascolari, la 16 cause esterne, la 11, altre malattie respiratorie non acute, la 4, malattie endocrine, nutrizionali o metaboliche e la 6, malattie del sistema nervoso e degli organi di senso.

Per le donne invece la situazione è diversa, le cause 6 e 11 non sono presenti in questo *range*, e anche l'ordine è mutato rispetto agli uomini. Ci sono infatti, sempre in ordine decrescente, dopo le cause già descritte e chiaramente visibili nel grafico complessivo relativo al 2000, la causa 8, malattie cerebrovascolari, la 2, neoplasma (per gli uomini questa causa ha un valore molto più elevato), la 10, malattie respiratorie acute, la 16, cause esterne, la 4, malattie endocrine e la 5, disturbi mentali e comportamentali, che nel modello per gli uomini non è considerata.

Il grafico relativo al 2006, il 4.15, non è di molto differente dal precedente se non per la stima della proporzione di decessi per la causa 7, per la classe 95 anni e oltre, i decessi per malattie cardiache infatti sono notevolmente più bassi del precedente: per il 2006, infatti, tale proporzione si aggira attorno all'11%, quindi circa 5 punti percentuali in meno dell'altro anno considerato. Per le donne tale valore resta attorno all'11% come per il 2000. Anche in questo caso per gli uomini la seconda causa per proporzione stimata è la 2 (neoplasma) e a seguire la 17, cioè altre cause o cause ignote, mentre per le donne c'è prima la 17. Anche per il 2006 è stato creato il grafico con maggior dettaglio per valori bassi delle proporzioni, il 4.16.

Per gli uomini la crescita delle due cause maggiori ossia la 7 (malattie cardiache) e la 2 (neoplasma) non mostra particolari differenze rispetto al 2000, mentre le altre cause manifestano nel complesso valori leggermente inferiori soprattutto per età verso l'estremo superiore; le cause con una porporzione inferiore a 2% nel 2006 per la classe 95 e oltre sono infatti 9 rispetto alle 7 del 2000). Anche per le donne si osserva questo fenomeno complessivo, le cause per l'ultima classe di età considerata con una stima della proporzione inferiore al 2% sono 11 nel 2006 e 9 nel 2000.

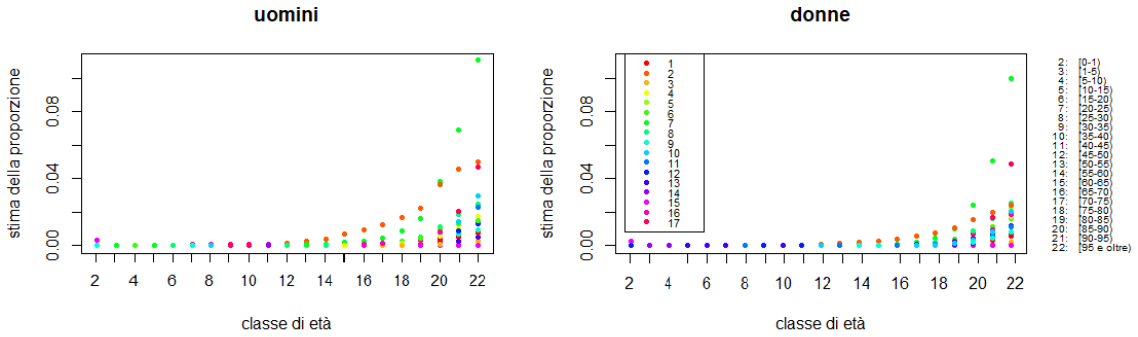

Figura 4.15: Stima della proporzione di decessi per classe di età, causa e sesso: anno 2006

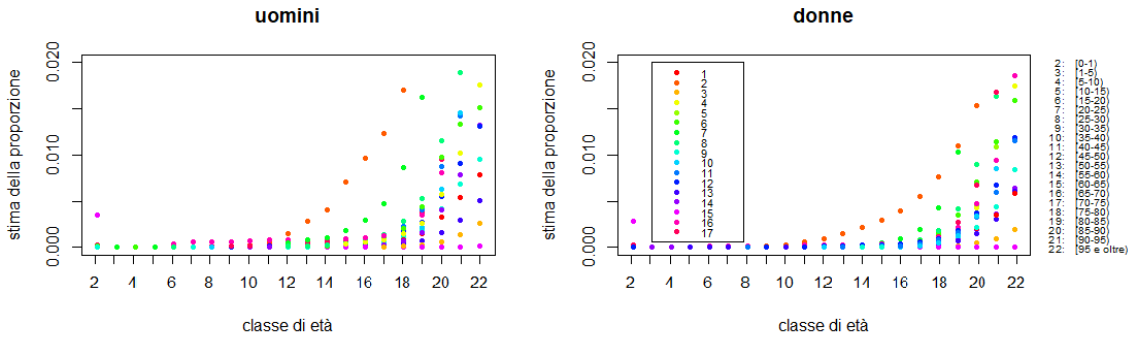

Figura 4.16: Approfondimento per valori bassi della stima della proporzione di decessi per classe di età, causa e sesso: anno 2006

Per quanto riguarda il 2013, le cui stime delle proporzioni sono riportate nel grafico 4.17, non sembrano esserci grosse differenze rispetto al 2006, né per gli uomini né per le donne. Ancora una volta la causa con proporzione stimata maggiore è la 7, malattie cardiache, sia per gli uomini che per le donne, anche se i primi hanno un valore leggermente più basso rispetto al precedente. A seguire ci sono per gli uomini la causa 2, neoplasma, e la 17, altre malattie o causa ignota; mentre per le donne la 17.

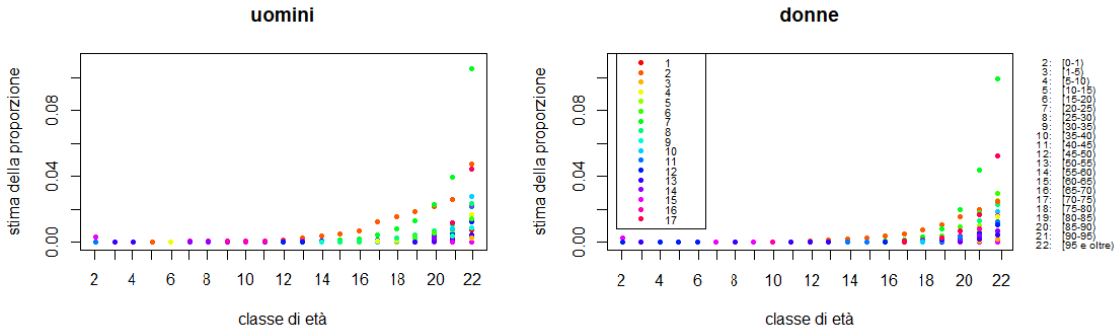

Figura 4.17: Stima della proporzione di decessi per classe di età, causa e sesso: anno 2013

Anche considerando il grafico 4.18 non si notano particolari stravolgimenti rispetto alle situazioni già descritte, per gli uomini la causa 2 tende ad avere valori stimati più bassi rispetto al 2006, mentre la 7 proporzioni più alte, soprattutto intorno agli 80 anni in su. Per le donne il discorso è analogo. Anche dal confronto tra i grafici dei due sessi non ci sono rilevanti differenze oltre a quelle già menzionate per gli altri anni.

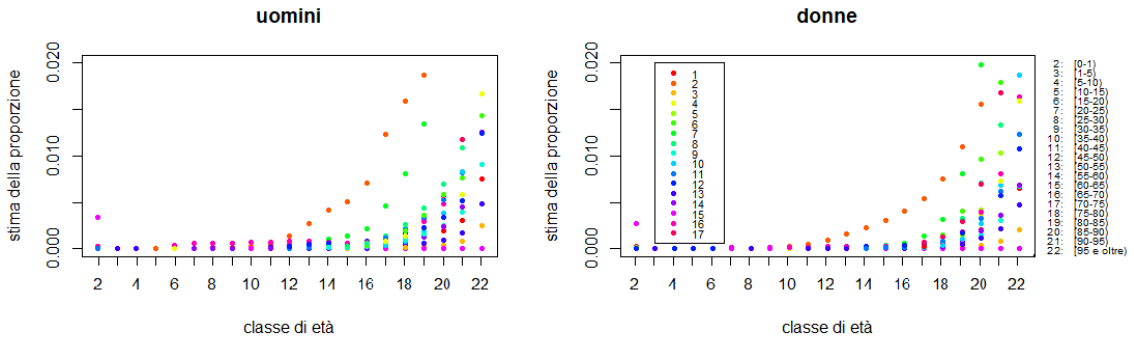

Figura 4.18: Approfondimento per valori bassi della stima della proporzione di decessi per classe di età, causa e sesso: anno 2013

Facendo quindi una rapida sintesi delle influenze che ha ogni covariata sulla stima della proporzione di decessi per causa si può notare che:

- l'anno non ha una particolare influenza né per gli uomini né per le donne, in entrambi i casi ha un coefficiente negativo che quindi porta ad una riduzione della stima al passare del tempo, come confermato dai grafici. A supporto di questa conclusione vi è la presenza nel 2000[4] di proporzioni di decessi stimate per gli uomini, in particolare nella causa 7 (malattie cardiache), molto più alte che per gli altri anni, fatto che per le donne è molto meno marcato: nel modello ZINB (senza la causa 5 per gli uomini) l'influenza dell'anno è più importante rispetto al GLM (con risposta Binomiale Negativa per le donne);
- la classe di età ha un'influenza più marcata per gli uomini rispetto alle donne, le proporzioni di decessi stimate per i primi aumentano di più rispetto alle donne al crescere dell'età degli individui;
- le cause hanno influenza diversa sulla stima della proporzione di decessi distintamente per uomini e donne, come si può vedere dai grafici appena descritti e dalle considerazione appena fatte. In particolare le donne hanno proporzioni generalmente più basse degli uomini e alcune cause (come la 2, la 6 e la 11, rispettivamente neoplasma, malattie del sistema nervoso e degli organi di senso e altre malattie respiratorie non acute diverse da quelle considerate nella causa 10) hanno un impatto di gran lunga inferiore rispetto a quello che hanno sugli uomini.

[4]Si ricorda che la variabile anno è stata riscalata per il minimo, ossia 2000, quindi per questo anno il coefficiente relativo viene moltiplicato per 0.

Capitolo 5

Il modello gerarchico

In questo capitolo verrà presentata un'ulteriore estensione dei modelli descritti nel capitolo 4, per cercare di interpretare al meglio i dati relativi alla mortalità per causa. Il capitolo viene suddiviso in due parti: nella prima ci sarà una breve descrizione teorica del modello gerarchico implementato, mentre nella seconda si vedrà l'applicazione del modello ai dati.

Tale estensione riguarda lo ZINB, in particolare il modello *Negative Binomial Zero Inflated* gerarchico, ossia un modello Binomiale Negativo con inflazione di zeri al quale viene stimata un'intercetta casuale diversa per ogni causa. Si è scelto di sviluppare questa estensione per la sola distribuzione Binomiale Negativa perchè, come emerso più volte nel capitolo 4, è quella che mostra un adattamento migliore ai dati rispetto alla distribuzione di Poisson.

5.1 La teoria

La scelta di implementare un modello di tipo gerarchico è dovuta al fatto che si può supporre che vi sia una struttura gerarchica all'interno dei dati, in particolare ci si potrebbe aspettare che tutti coloro cha sono deceduti per una determinata causa abbiano un qualcosa in comune. Il modello *Multilevel*, ad intercetta casuale, è un modello per dati gerarchici in cui solo l'intercetta varia tra i gruppi [Gelman e Hill 2007]. Innanzitutto pare doveroso distinguere tra unità di primo livello e di secondo

livello, in particolare queste ultime sono considerate le cause mentre le prime gli individui deceduti.

La scelta di usare un modello *Multilevel* è un compromesso tra un modello *pooling* e un *no pooling*, in particolare:

- modello *pooling*: si suppone che i dati facciano parte di un unico campione (come per i modelli presenti nel capitolo 4), l'adattamento ai dati è ridotto e vi è un'assenza di considerazione della variabilità tra le unità di secondo livello;
- modello *no pooling*: si suppone che i dati provengano da tanti gruppi quante sono le unità di secondo livello, l'adattamento ai dati è fin troppo eccessivo e la variabilità tra i gruppi tende ad essere sovrastimata;
- modello *Multilevel*: ogni unità di secondo livello ha una sua intercetta, mentre gli altri coefficienti sono uguali, questo modello permette di utilizzare una via di mezzo tra i due precedenti in modo da sfruttarne le opportunità di ciascuno e sopperire alle inadeguatezze.

L'equazione del modello ZINB ad intercetta casuale è quindi del tipo:

$$\mathbb{P}(Y=y)\begin{cases}(1-\pi_i)\binom{y}{k}\dfrac{\rho^{\kappa}}{(1-\rho)^{\kappa}}(1-\rho)^{y} & \text{per } y=0 \text{ e } i=1,...,17;\\ \pi_i+(1-\pi_i)\cdot\rho^{\kappa} & \text{per } y=0 \text{ e } i=1,...,17.\end{cases}$$

Per quanto riguarda la stima dei parametri del modello specifico, questa è possibile tramite una procedura numerica di massimizzazione della log-verosimiglianza penalizzata attraverso l'algoritmo EM che è una procedura di massimizzazione[*Estimation of parameters by multi-level zero inflated Poisson and Zero Inflated Negative Binomial Regression models*]. Questo algoritmo, chiamato di *Expectation Maximization*, viene usato generalmente nel caso di dati mancanti ed è una procedura iterativa composta da due passi, si alterna una massimizzazione condizionata a un'estensione dell'algoritmo di Newton Raphson, già spiegato nel capitolo 3.

5.2 Applicazione ai dati

Si è deciso di implementare il modello descritto nella sezione precedente per i soli dati relativi agli uomini, questo perchè già nel capitolo 4 è stato più volte ribadito come per la parte di osservazioni relativa alle donne, usare un modello ad inflazione di zeri, non fosse una scelta adeguata, il coefficiente relativo infatti sia per lo ZIP che lo ZINB erano non significativamente diversi da 0. Inoltre per i dati relativi agli uomini vi è la presenza costante di una causa, la 5, relativa ai disturbi mentali e comportamentali, che non risulta mai significativa, un modello come quello usato in questo capitolo, potrebbe permettere di cogliere maggiormente le dinamiche che provocano tali risultati.

Per stimare il modello ZINB *Multilevel* è stato usato ancora una volta il *software* *R*, ma è necessario scaricare la libreria *glmmTMB* [Simpson 2017a, AAVV 2017] e il comando da usare prende lo stesso nome: *glmmTMB{....family = nbinom1(link = "log")}*[1]. Come si può notare dagli argomenti del comando è stata scelta ancora una volta come funzione legame il logaritmo, per completezza rispetto ai modelli precedentemente stimati.

Nella tabella 5.1 sono riportati i coefficienti relativi alle covariate del modello ZINB con le usuali misure utili per la loro interpretazione.

	Uomini		
	Stima	*Std Error*	*p-value*
intercetta	-9.1543	0.2454	0.0000
$[1-5)$	-1.3331	0.0914	0.0000
$[5-10)$	-1.7286	0.0930	0.0000
$[10-15)$	-1.6608	0.0923	0.0000
$[15-20)$	-1.3158	0.0886	0.0000
$[20-25)$	-1.0620	0.0864	0.0000
$[25-30)$	-0.8860	0.0851	0.0000

[1]La distribuzione scelta, ossia *nbinom1* è dovuta alla possibilità che questa fornisce di modellare situazioni in cui la sovra-dispersione non sia così elevata come permetterebbe di modellare *nbinom2* [Bolker 2016]. Nel modello ZINB era stato osservato che utilizzare la classica distribuzione Binomiale Negativa portava a considerare una sovra-dispersione troppo alta.

$[30-35)$	-0.7039	0.0834	0.0000
$[35-40)$	-0.4007	0.0810	0.0000
$[40-45)$	0.0280	0.0780	0.720
$[45-50)$	0.4972	0.0752	0.0000
$[50-55)$	0.9006	0.0732	0.0000
$[56-60)$	1.2467	0.0722	0.0000
$[60-65)$	1.5684	0.0718	0.0000
$[65-70)$	1.9261	0.0713	0.0000
$[70-75)$	2.3502	0.0704	0.0000
$[75-80)$	2.8392	0.0697	0.0000
$[80-85)$	3.3818	0.0694	0.0000
$[85-90)$	3.9500	0.0697	0.0000
$90-95)$	4.4883	0.0711	0.0000
[95 e oltre)	5.0843	0.0756	0.0000
anno	-0.0194	0.0016	0.0000
intercetta	-9.7320	1.8970	0.0000

Tabella 5.1: Stima dei parametri del modello ZINB. La prima parte (prima della riga) è riferita alla distribuzione Binomiale Negativa mentra la seconda alla Binomiale*Multilevel*

Dalla tabella relativa ai coefficienti delle covariate si può notare come l'unico parametro non significativo sia quello relativo alla classe di età $[40-45)$.

Il parametro di sovra-disperisione stimato è pari a 193, un valore così elevato fa intuire che la scelta della distribuzione Binomiale Negativa, anzichè la Poisson, sia piuttosto sensata.

Nella tabella 5.2 sono riportate le intercette casuali delle 17 cause considerate, come si può notare alcune sono positive e altre negative, ma la maggior parte sono intorno allo 0: sintomo che non vi sono comportamenti molto diversi per cause differenti. Questi valori, infatti, vanno contemplati assieme alle altre due intercette stimate nel modello, ossia quella relativa alla distribuzione Binomiale Negativa e quella della parte *Zero Inflated*. Valori così bassi delle intercette relative alle cause portano a differenze piuttosto contenute tra i modelli stimati (si ricorda infatti che

causa	intercetta
causa 1	-0.2929
causa 2	2.1887
causa 3	-1.3085
causa 4	0.0545
causa 5	-0.1204
causa 6	0.2769
causa 7	1.5849
causa 8	0.4184
causa 9	-0.2927
causa 10	-0.3279
causa 11	0.2066
causa 12	0.3482
causa 13	-1.1111
causa 14	-0.7145
causa 15	-1.8275
causa 16	0.7379
causa 17	0.6097

Tabella 5.2: Intercette casuali per le cause del modello ZINB *Multilevel*

solo le intercette sono diverse, gli altri coefficienti coincidono e sono quelli riportati in tabella 5.1).

La prima analisi diagnostica, presente nel grafico 5.1, è quella che riporta la stima della proporzione di decessi contro il valore osservato.

Il grafico non illustra una buona situazione: i punti all'aumentare della proporzione tendono ad allontanarsi sempre più dalla retta bisettrice del primo quadrante, dove dovrebbero giacere in una situazione ideale. Ma dal grafico 5.1 non è chiaro se sia presente un andamento sottostante che non si riesce a cogliere. In aiuto di questo quesito sono riportati i grafici 5.2.

Da questi si può notare come le proporzioni che si discostano dai valori stimati siano ragruppate per causa, in particolare alcune sono adeguatamente colte dal mo-

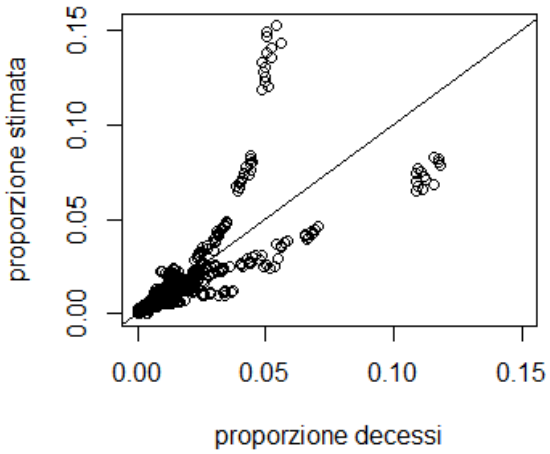

Figura 5.1: Stima della proporzione contro valore osservato del modello ZINB *Multilevel*: uomini

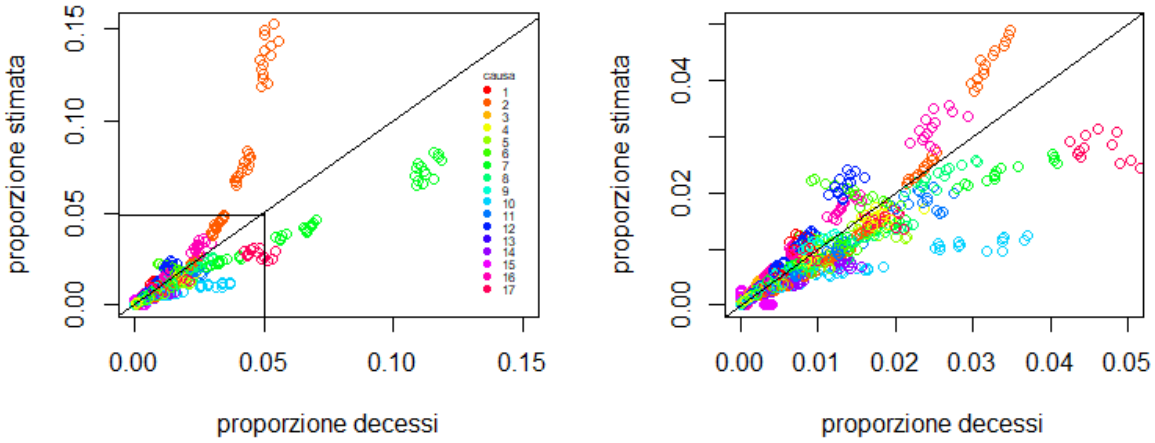

Figura 5.2: Stima della propozione contro valore osservato del modello ZINB *Multilevel* con colori diversi per causa e approfondimento per valori bassi (nel grafico a destra): uomini

dello ZINB *Multilevel* (soprattutto quelle che hanno una proporzione molto vicina allo 0) mentre altre si allontanano di molto dalla situazione ideale. In particolare approfondendo l'analisi dei più più estremi si osserva che il gruppo di colore verde che si discosta verso il basso, quindi per cui la proporzione osservata è più alta di quella stimata è rappresentato da individui morti per la causa 7, ossia le malattie cardiache, e il gruppo collocato più a destra sono gli individui che appartengono alla classe di età 95 anni e oltre. L'altro gruppo che si discosta in questo caso verso l'alto, quello arancione, rappresenta la causa 2, decessi per neoplasma, e lo stesso gruppo di età, ossia coloro che hanno dai 95 anni in su, in questo caso il modello sovrastima notevolmente la proporzione dei decessi rispetto a quella osservata. Dal grafico a destra, specifico per valori della proporzione di decessi vicini allo 0, si può notare come in generale ci siano delle cause che non sono colte bene dal modello. A conferma di questo si possono vedere anche i grafici 5.3.

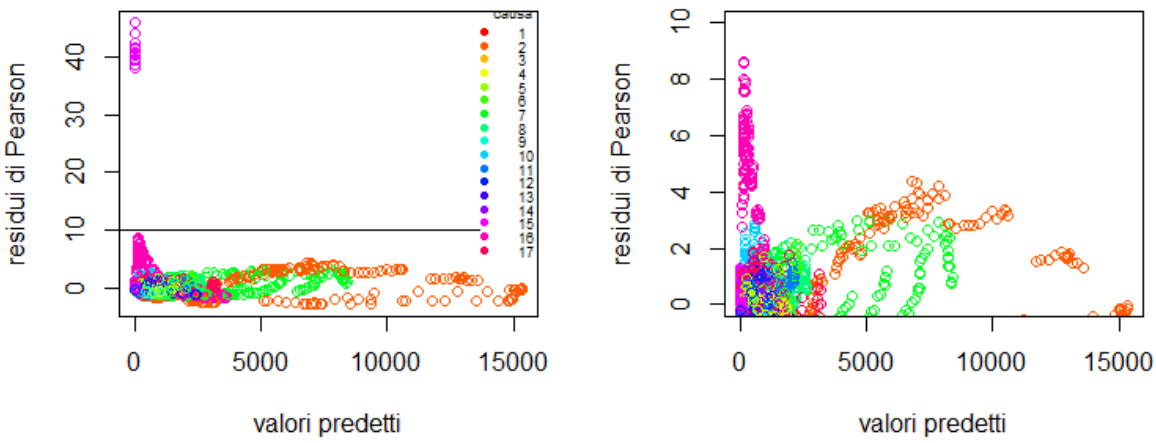

Figura 5.3: Residui di Pearson contro valori predetti del modello ZINB *Multilevel* con colori diversi per causa e approfondimento per valori bassi (nel grafico a destra): uomini

Questi riportano i valori predetti contro i residui di Pearson, si può notare, anche in questo caso, come i punti siano raggruppati per causa, e quindi per colore, e che seguano andamenti simili per la stessa. Un'accurata ispezione fa emergere che le cause con residuo di Pearson più elevato siano in ordine la 15, la 16, la

2 e la 7, mentre le classi di età che creano maggiori problemi sono *in primis* la classe relativa ai neonati, che non hanno ancora compiuto il primo anno di età (non sembra casuale che gli unici individui deceduti per la causa 15 siano appunto gli appartenenti a questa classe di età) e a seguire le classi dai 20 ai 40 anni circa. In effetti i decessi in questi intervalli, come dimostrato dalle analisi esplorative al capitolo 2 sono veramente pochi. A confermare che queste fossero le tre cause con maggiori particolarità vi era anche l'analisi della significatività dei coefficienti nel modello ZINB (non gerarchico) scelto nella sezione 4.3.1 del capitolo precedente, per queste tre cause infatti i parametri relativi alle interazioni con le classi di età erano tutti diversi da 0.

Per quanto riguarda l'anno di rilevazione invece non sembrano emergere particolari criticità.

Un'ultima considerazione è quella che riguarda il valore del criterio di AIC che è di circa 62015, che confrontato con quello relativo agli altri modelli presentati nel capitolo 4 è di gran lunga superiore (il minore era riferito al modello GLM con risposta Binomiale Negativa ed era attorno a 41965), ma in questo modello sono stati inseriti i soli effetti principali a differenza degli altri dove erano presenti anche le interazioni di primo grado.

5.3 Criticità e potenzialità del modello

Da quanto emerso nelle analisi diagnostiche il modello gerarchico implementato non è del tutto soddisfacente, potrebbe essere un buon punto di partenza per implementare un modello più adatto ai dati di cui si dispone ma certamente con delle modifiche.

I grafici diagnostici fanno intuire che un modello ad intercetta casuale sulle cause potrebbe non essere sufficiente per carpire le dinamiche intrinseche ai dati usati. Potrebbe essere utile inserire un'intercetta casuale anche sulle classi di età, quindi sviluppare un modello su 3 livelli, oppure inserire degli effetti casuali anche per le covariate. L'uso del modello ad intercetta casuale sulle cause provoca una sorta di *partial pooling*, una *mean srinkage*, ossia una compressione delle stime verso la media, questo può provocare delle forzature, se come in questo caso, vi sono situazioni

particolarmente estreme, come lo sono le cause 2, 7, 15 e 16 (che sono rispettivamente 2 neoplasma, 7 malattie cardiache, 15 condizioni originarie del periodo perinatale e anomalie o malformazioni congenite e 16 cause esterne) in relazione alle diverse classi di età. Queste stime vengono compresse verso la media ma non sempre questa è una buona strategia per ottimizzare i dati disponibili. Il fenomeno sottostante che riguarda le diverse cause è profondamente diverso se si considerano le diverse classi di età, la causa 15 ad esempio mostra una proporzione praticamente nulla per tutte le classi ad eccezione della prima, quella relativa ai neonati nel primo anno di vita: forzare un modello con una sola intercetta per tutte le classi di età preclude la possibilità di riuscire a cogliere questo fatto, a maggior ragione se per le altre cause, in particolare per la 2, la 7 e la 16 che presentano una situazione opposta, cioè bassissimi valori, praticamente nulli, per classi medio basse, che diventano piuttosto importanti quando ci si sposta verso l'estremo superiore dell'età. Ma tutto questo è intrinseco alla natura stessa delle varie cause e un modello *Multilevel*, come quello stimato, non riesce evidentemente a cogliere avendo dei coefficienti fissi per le varie covariate, in particolare alle classi di età.

Un altro problema, di certo non irrilevante, è dato dall'impossibilità di stimare un modello più complesso, che potrebbe in parte risolvere il limite appena discusso. Avendo a disposizione altre informazioni sui gruppi di individui, come potrebbero essere l'uso di alcool, fumo o altre abitudini che si suppongano essere influenti sullo stato di salute, la posizione sociale o il reddito, si potrebbe costruire un modello più complesso che catturi maggiormente le dinamiche intrinseche ai dati.

Conclusioni e possibili sviluppi

Dalle analisi riportate nel presente elaborato è emerso, senza dubbio, come, per questi specifici dati, la distribuzione Binomiale Negativa si adatti meglio alla distribuzione di Poisson, indipendentemente dal tipo di modello statistico utilizzato.

Questo è spiegato dalla presenza di sovradispersione nei dati; la distribuzione di Poisson non riesce a coglierla, infatti per tale distribuzione la media e la varianza sono assunte uguali. Grazie alla distribuzione Binomiale Negativa è invece possibile cogliere questa caratteristica, a conferma di ciò si può notare che il parametro di dispersione ($log(\theta)$) nei vari modelli, una volta riportato nella scala originale, è sempre piuttosto elevato (sintomo che vi è la presenza di sovradispersione che il modello di Poisson non riesce a cogliere).

I dati relativi alle donne e agli uomini sono differenti, la proporzione di decessi per causa è infatti influenzata da fattori diversi, questo è stato più volte ribadito in letteratura e confermato, anche in questo caso, dai dati disponibili. Si sono resi necessari due modelli differenti per cogliere le dinamiche intrinseche ai dati e per cercare di stimare al meglio le proporzioni, in particolare per le donne l'utilizzo di un modello *Zero Inflated* non è adeguato, mentre per gli uomini sì. Inoltre la causa 5, relativa a disturbi mentali e comportamentali, ha causato non pochi problemi, tant'è che si è deciso di rimuoverla dal *dataset* relativo agli uomini. Una possibile estensione di questo lavoro potrebbe essere indirizzata verso un approfondimento di queste problematicità, al fine di scoprire quali siano i fattori che rendono questa causa così particolare, quantomeno per gli uomini.

Potrebbe essere interessante anche riproporre le stesse analisi o condurne di differenti utilizzando come classificazione delle cause, la *Intermediate List* anzichè la *Short*, come è stato fatto in questo lavoro, e in quel caso, aumentando gli zeri, po-

trebbe risultare utile un modello *Zero Inflated* anche per le donne e per gli uomini sarebbe forse possibile introdurre più covariate nella parte relativa alla distribuzione Binomiale. Un'altra possibile estensione potrebbe essere di tipo geografico, ossia riproporre le stesse analisi per altri Stati, europei o non, e osservare se vi siano differenze interessanti rispetto alla Francia.

Per gli uomini è stato, inoltre, implementato anche un modello ZINB di tipo gerarchico, ad intercetta casuale: questo modello ha dei limiti anche se appare una strada promettente nel caso si avessero a disposizione maggiori informazioni relative alle caratteristiche degli individui considerati. Una possibile estensione potrebbe essere da una parte quella di usare più covariate in modo da individuare meglio i fattori che influenzano la proporzione di decessi, e dall'altre introdurre effetti casuali anche sulle classi di età, ad esempio.

Bibliografia

AAVV (2017). *Package 'glmmTMB', Generalized Linear Mixed Models using Template Model Builder*. https://cran.r-project.org/web/packages/glmmTMB/glmmTMB.pdf. data downloaded: 14-10-2017.

Azzalini, Adelchi (2001). *Inferenza Statistica: una rappresentazione basata sul concetto di verosimiglianza, 2a edizione*. Milano: Springer.

Azzalini, Adelchi e Bruno Scarpa (2012). *Data Analysis and Data Mining: An Introduction*. OUP USA.

Bolker, Ben (2016). *Getting started with the glmmTMB package*. https://cran.r-project.org/web/packages/glmmTMB/vignettes/glmmTMB.pdf. data downloaded: 14-10-2017.

Demographic Studies (France), French Institute for e Max Planck Institute for Demographic Research (Germany). *Human Cause-of-Death Database*. http://www.causeofdeath.org.

Estimation of parameters by multi-level zero inflated Poisson and Zero Inflated Negative Binomial Regression models. http://shodhganga.inflibnet.ac.in/bitstream/10603/96838/14/14_chapter206.pdf. data downloaded: 10-10-2017.

Gelman, Andrew e Jeniffer Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Jorgensen, Bent. *Generalized linear models*. https://www.ime.usp.br/~abe/lista/pdftGzimaFtH4.pdf. data downloaded: 06-10-2017.

Missov, Campos De Lima Lanza Queiroz e Lenart (2016). «Methods to Estimate Mortality Curves in Small Areas: an Application to Municipality Data in Brazil». In:

Nelder, J.A. e R.W.M. Wedderburn (1972). «Generalized linear models». In: *Journal of the Royal Statistical Society Series A*.

Pace, Luigi e Alessandra Salvan (2001). *Introduzione alla statistica: Inferenza, verosimiglianza, modelli*. Verona: CEDAM.

Simpson, Gavin (2017a). *Fitting count and zero-inflated count GLMMs with mgcv*. `http://www.fromthebottomoftheheap.net/2017/05/04/compare-mgcv-with-glmmTMB/`. data downloaded: 12-10-2017.

— (2017b). *Rilevazioni sui decessi e sulle cause di morte: informazioni sulla rilevazione*. `https://www.istat.it/it/archivio/4216`. data downloaded: 24-10-2017.

Usai, Federica (2011). *Epidemiologia delle strongilosi dell'asino: quali applicazioni per il controllo delle infezioni da elminti?* Tesi di Dottorato presso Alma Mater Studiorum-Università di Bologna.

Viviano, Lorena (2008). «Modelli di conteggio con eccesso di zeri: due appprocci a confronto». In: *Department of Information Technology and Mathematical Methods Working Paper*.

Appendice A

In questa parte saranno inseriti gli *Outputt* dei modelli stimati descritti nel resto dell'elaborato, ci si limita ad inserire le interazioni di primo grado per evitare di produrre una sezione troppo lunga e dispersiva.

A.1 Il modello ZINB *Multilevel* per gli uomini

	Stima	*Std.Error*	*p-value*
causa 1	3.5154	0.0650	0.0000
causa 2	2.6351	0.0857	0.0000
causa 3	2.2607	0.0987	0.0000
causa 4	3.6458	0.0630	0.0000
causa 6	4.1537	0.0566	0.0000
causa 7	3.2255	0.0705	0.0000
causa 8	1.7097	0.1240	0.0000
causa 9	0.4381	0.2229	0.0493
causa 10	2.4237	0.0926	0.0000
causa 11	1.7117	0.1244	0.0000
causa 12	2.7581	0.0866	0.0000
causa 13	-1.2202	0.5021	0.0151
causa 14	0.7603	0.1912	0.0001
causa 15	7.2633	0.0458	0.0000
causa 16	3.9386	0.0590	0.0000
causa 17	4.7397	0.0520	0.0000
$[1-5)$	-0.4019	0.0975	0.0000
$[5-10)$	-1.8313	0.1412	0.0000
$[10-15)$	-1.7649	0.1380	0.0000
$[15-20)$	-1.0914	0.1130	0.0000
$[20-25)$	-0.8750	0.1072	0.0000

$[25-30)$	-0.2690	0.0954	0.0048
$[30-35)$	0.6089	0.0861	0.0000
$[35-40)$	1.4938	0.0820	0.0000
$[40-45)$	1.9322	0.0809	0.0000
$[45-50)$	2.0349	0.0808	0.0000
$[50-55)$	2.0793	0.0807	0.0000
$[55-60)$	2.1432	0.0806	0.0000
$[60-65)$	2.2621	0.0804	0.0000
$[65-70)$	2.4086	0.0802	0.0000
$[70-75)$	2.7469	0.0798	0.0000
$[75-80)$	3.1605	0.0796	0.0000
$[80-85)$	3.3517	0.0795	0.0000
$[85-90)$	3.2085	0.0795	0.0000
$[90-95)$	2.6435	0.0800	0.0000
[95 e oltre)	1.5804	0.0818	0.0000
anno	-0.0051	0.0008	0.0000
causa 2 : $[1-5)$	1.8188	0.1419	0.0000
causa 3 : $[1-5)$	0.7658	0.1631	0.0000
causa 4 : $[1-5)$	-0.1874	0.1377	0.1736
causa 6 : $[1-5)$	-0.2336	0.1300	0.0724
causa 7 : $[1-5)$	-0.3421	0.1505	0.0230
causa 8 : $[1-5)$	-0.1068	0.2217	0.6299
causa 9 : $[1-5)$	-0.6963	0.4517	0.1232
causa 10 : $[1-5)$	-0.2091	0.1793	0.2435
causa 11 : $[1-5)$	0.3874	0.2012	0.0542
causa 12 : $[1-5)$	-0.4336	0.1743	0.0128
causa 13 : $[1-5)$	1.5806	0.5836	0.0068
causa 14 : $[1-5)$	-0.4787	0.3625	0.1866
causa 15 : $[1-5)$	-3.0793	0.1237	0.0000
causa 16 : $[1-5)$	1.3091	0.1248	0.0000
causa 17 : $[1-5)$	-0.6412	0.1267	0.0000
causa 2 : $[5-10)$	3.3676	0.1744	0.0000
causa 3 : $[5-10)$	1.2409	0.2134	0.0000
causa 4 : $[5-10)$	0.5209	0.1818	0.0042
causa 6 : $[5-10)$	0.6010	0.1706	0.0004
causa 7 : $[5-10)$	0.3433	0.1997	0.0857
causa 8 : $[5-10)$	0.7340	0.2781	0.0083
causa 9 : $[5-10)$	0.1728	0.5673	0.7606
causa 10 : $[5-10)$	-0.1070	0.2757	0.6980
causa 11 : $[5-10)$	0.8458	0.2704	0.0018
causa 12 : $[5-10)$	-0.0063	0.2426	0.9792
causa 13 : $[5-10)$	3.0105	0.5925	0.0000
causa 14 : $[5-10)$	-0.4360	0.6260	0.4860
causa 15 : $[5-10)$	-2.7307	0.1702	0.0000
causa 16 : $[5-10)$	2.2218	0.1624	0.0000
causa 17 : $[5-10)$	-0.2831	0.1731	0.1020
causa 2 : $[10-15)$	3.1967	0.1722	0.0000
causa 3 : $[10-15)$	0.9932	0.2181	0.0000

causa 4 : [10 − 15)	0.6002	0.1766	0.0007
causa 6 : [10 − 15)	0.8064	0.1652	0.0000
causa 7 : [10 − 15)	0.8298	0.1830	0.0000
causa 8 : [10 − 15)	1.3809	0.2367	0.0000
causa 9 : [10 − 15)	1.2858	0.3843	0.0008
causa 10 : [10 − 15)	0.0317	0.2583	0.9022
causa 11 : [10 − 15)	1.4899	0.2325	0.0000
causa 12 : [10 − 15)	0.2302	0.2244	0.3050
causa 13 : [10 − 15)	3.2123	0.5762	0.0000
causa 14 : [10 − 15)	-0.2139	0.5544	0.6997
causa 15 : [10 − 15)	-2.6759	0.1660	0.0000
causa 16 : [10 − 15)	2.6391	0.1585	0.0000
causa 17 : [10 − 15)	-0.0996	0.1671	0.5511
causa 2 : [15 − 20)	3.0498	0.1512	0.0000
causa 3 : [15 − 20)	0.5940	0.1928	0.0021
causa 4 : [15 − 20)	0.1124	0.1545	0.4669
causa 6 : [15 − 20)	0.8726	0.1393	0.0000
causa 7 : [15 − 20)	1.0607	0.1509	0.0000
causa 8 : [15 − 20)	1.4390	0.1990	0.0000
causa 9 : [15 − 20)	1.8081	0.2962	0.0000
causa 10 : [15 − 20)	-0.3542	0.2263	0.1175
causa 11 : [15 − 20)	1.6403	0.1948	0.0000
causa 12 : [15 − 20)	0.1342	0.1857	0.4700
causa 13 : [15 − 20)	3.0383	0.5501	0.0000
causa 14 : [15 − 20)	-0.3289	0.4407	0.4555
causa 15 : [15 − 20)	-3.1232	0.1427	0.0000
causa 16 : [15 − 20)	3.7963	0.1355	0.0000
causa 17 : [15 − 20)	0.7907	0.1354	0.0000
causa 2 : [20 − 25)	3.1744	0.1462	0.0000
causa 3 : [20 − 25)	0.4740	0.1864	0.0110
causa 4 : [20 − 25)	0.1683	0.1462	0.2497
causa 6 : [20 − 25)	0.8759	0.1336	0.0000
causa 7 : [20 − 25)	1.2952	0.1428	0.0000
causa 8 : [20 − 25)	1.9326	0.1831	0.0000
causa 9 : [20 − 25)	2.1503	0.2770	0.0000
causa 10 : [20 − 25)	0.3564	0.1819	0.0500
causa 11 : [20 − 25)	1.8037	0.1850	0.0000
causa 12 : [20 − 25)	0.5675	0.1655	0.0006
causa 13 : [20 − 25)	2.8561	0.5478	0.0000
causa 14 : [20 − 25)	0.5042	0.3164	0.1110
causa 15 : [20 − 25)	-3.2360	0.1369	0.0000
causa 16 : [20 − 25)	4.1052	0.1305	0.0000
causa 17 : [20 − 25)	1.3160	0.1289	0.0000
causa 2 : [25 − 30)	2.8587	0.1372	0.0000
causa 3 : [25 − 30)	-0.1431	0.1802	0.4271
causa 4 : [25 − 30)	-0.1958	0.1347	0.1459
causa 6 : [25 − 30)	0.3277	0.1241	0.0083
causa 7 : [25 − 30)	1.1928	0.1313	0.0000

causa 8 : [25 − 30)	1.6220	0.1730	0.0000
causa 9 : [25 − 30)	1.9866	0.2633	0.0000
causa 10 : [25 − 30)	0.0275	0.1674	0.8693
causa 11 : [25 − 30)	1.3623	0.1762	0.0000
causa 12 : [25 − 30)	0.6883	0.1473	0.0000
causa 13 : [25 − 30)	2.6208	0.5358	0.0000
causa 14 : [25 − 30)	0.9625	0.2547	0.0002
causa 15 : [25 − 30)	-3.8578	0.1281	0.0000
causa 16 : [25 − 30)	3.4788	0.1210	0.0000
causa 17 : [25 − 30)	0.9734	0.1188	0.0000
causa 2 : [30 − 35)	2.4162	0.1304	0.0000
causa 3 : [30 − 35)	-0.6561	0.1650	0.0001
causa 4 : [30 − 35)	-0.8279	0.1257	0.0000
causa 6 : [30 − 35)	-0.2949	0.1161	0.0111
causa 7 : [30 − 35)	1.0235	0.1223	0.0000
causa 8 : [30 − 35)	1.3086	0.1636	0.0000
causa 9 : [30 − 35)	1.7874	0.2518	0.0000
causa 10 : [30 − 35)	-0.2137	0.1497	0.1534
causa 11 : [30 − 35)	0.7349	0.1684	0.0000
causa 12 : [30 − 35)	0.7015	0.1349	0.0000
causa 13 : [30 − 35)	2.2097	0.5258	0.0000
causa 14 : [30 − 35)	0.3556	0.2432	0.1437
causa 15 : [30 − 35)	-4.5853	0.1195	0.0000
causa 16 : [30 − 35)	2.6566	0.1138	0.0000
causa 17 : [30 − 35)	0.3690	0.1112	0.0009
causa 2 : [35 − 40)	2.2037	0.1273	0.0000
causa 3 : [35 − 40)	-1.2315	0.1561	0.0000
causa 4 : [35 − 40)	-1.0958	0.1183	0.0000
causa 6 : [35 − 40)	-0.8066	0.1119	0.0000
causa 7 : [35 − 40)	0.9590	0.1181	0.0000
causa 8 : [35 − 40)	1.0638	0.1586	0.0000
causa 9 : [35 − 40)	1.4955	0.2466	0.0000
causa 10 : [35 − 40)	-0.6944	0.1423	0.0000
causa 11 : [35 − 40)	0.4043	0.1618	0.0125
causa 12 : [35 − 40)	0.8843	0.1291	0.0000
causa 13 : [35 − 40)	1.7544	0.5202	0.0007
causa 14 : [35 − 40)	-0.0568	0.2312	0.8059
causa 15 : [35 − 40)	-5.3868	0.1156	0.0000
causa 16 : [35 − 40)	1.9420	0.1107	0.0000
causa 17 : [35 − 40)	-0.1523	0.1076	0.1572
causa 2 : [40 − 45)	2.6218	0.1263	0.0000
causa 3 : [40 − 45)	-1.3446	0.1502	0.0000
causa 4 : [40 − 45)	-0.9711	0.1152	0.0000
causa 6 : [40 − 45)	-0.8748	0.1103	0.0000
causa 7 : [40 − 45)	1.2467	0.1168	0.0000
causa 8 : [40 − 45)	1.2760	0.1564	0.0000
causa 9 : [40 − 45)	1.7167	0.2439	0.0000
causa 10 : [40 − 45)	-0.6460	0.1376	0.0000

causa 11 : [40 − 45)	0.5043	0.1587	0.0015
causa 12 : [40 − 45)	1.3391	0.1274	0.0000
causa 13 : [40 − 45)	1.7752	0.5165	0.0006
causa 14 : [40 − 45)	0.1607	0.2221	0.4694
causa 15 : [40 − 45)	-5.9148	0.1159	0.0000
causa 16 : [40 − 45)	1.6224	0.1099	0.0000
causa 17 : [40 − 45)	-0.2675	0.1066	0.0121
causa 2 : [45 − 50)	3.3858	0.1260	0.0000
causa 3 : [45 − 50)	-1.1250	0.1461	0.0000
causa 4 : [45 − 50)	-0.5086	0.1136	0.0000
causa 6 : [45 − 50)	-0.6916	0.1097	0.0000
causa 7 : [45 − 50)	1.6925	0.1165	0.0000
causa 8 : [45 − 50)	1.7032	0.1556	0.0000
causa 9 : [45 − 50)	2.2234	0.2426	0.0000
causa 10 : [45 − 50)	-0.2568	0.1349	0.0570
causa 11 : [45 − 50)	1.0235	0.1569	0.0000
causa 12 : [45 − 50)	1.8301	0.1270	0.0000
causa 13 : [45 − 50)	2.2002	0.5141	0.0000
causa 14 : [45 − 50)	0.6707	0.2176	0.0021
causa 15 : [45 − 50)	-5.6535	0.1120	0.0000
causa 16 : [45 − 50)	1.5771	0.1097	0.0000
causa 17 : [45 − 50)	-0.0556	0.1063	0.6010
causa 2 : [50 − 55)	3.9889	0.1259	0.0000
causa 3 : [50 − 55)	-0.8132	0.1429	0.0000
causa 4 : [50 − 55)	-0.0601	0.1129	0.5943
causa 6 : [50 − 55)	-0.4713	0.1093	0.0000
causa 7 : [50 − 55)	2.1038	0.1163	0.0000
causa 8 : [50 − 55)	2.0968	0.1552	0.0000
causa 9 : [50 − 55)	2.6942	0.2420	0.0000
causa 10 : [50 − 55)	0.1695	0.1334	0.2036
causa 11 : [50 − 55)	1.6492	0.1559	0.0000
causa 12 : [50 − 55)	2.1817	0.1268	0.0000
causa 13 : [50 − 55)	2.6287	0.5127	0.0000
causa 14 : [50 − 55)	1.1167	0.2156	0.0000
causa 15 : [50 − 55)	-5.5674	0.1109	0.0000
causa 16 : [50 − 55)	1.4985	0.1097	0.0000
causa 17 : [50 − 55)	0.1057	0.1062	0.3194
causa 2 : [55 − 60)	4.2933	0.1259	0.0000
causa 3 : [55 − 60)	-0.6833	0.1415	0.0000
causa 4 : [55 − 60)	0.2597	0.1124	0.0209
causa 6 : [55 − 60)	-0.3185	0.1090	0.0035
causa 7 : [55 − 60)	2.3440	0.1162	0.0000
causa 8 : [55 − 60)	2.3426	0.1550	0.0000
causa 9 : [55 − 60)	3.0990	0.2417	0.0000
causa 10 : [55 − 60)	0.4277	0.1326	0.0013
causa 11 : [55 − 60)	2.0776	0.1554	0.0000
causa 12 : [55 − 60)	2.2812	0.1267	0.0000
causa 13 : [55 − 60)	3.0963	0.5118	0.0000

causa 14 : [55 − 60)	1.5958	0.2141	0.0000
causa 15 : [55 − 60)	-5.4280	0.1094	0.0000
causa 16 : [55 − 60)	1.2854	0.1096	0.0000
causa 17 : [55 − 60)	0.1834	0.1061	0.0838
causa 2 : [60 − 65)	4.3477	0.1257	0.0000
causa 3 : [60 − 65)	-0.5199	0.1398	0.0002
causa 4 : [60 − 65)	0.3822	0.1121	0.0007
causa 6 : [60 − 65)	-0.2407	0.1087	0.0268
causa 7 : [60 − 65)	2.4300	0.1160	0.0000
causa 8 : [60 − 65)	2.5046	0.1548	0.0000
causa 9 : [60 − 65)	3.2622	0.2415	0.0000
causa 10 : [60 − 65)	0.6119	0.1320	0.0000
causa 11 : [60 − 65)	2.3684	0.1551	0.0000
causa 12 : [60 − 65)	2.1953	0.1266	0.0000
causa 13 : [60 − 65)	3.2753	0.5115	0.0000
causa 14 : [60 − 65)	1.8761	0.2134	0.0000
causa 15 : [60 − 65)	-5.6627	0.1101	0.0000
causa 16 : [60 − 65)	0.9840	0.1096	0.0000
causa 17 : [60 − 65)	0.0640	0.1059	0.5455
causa 2 : [65 − 70)	4.3248	0.1256	0.0000
causa 3 : [65 − 70)	-0.4732	0.1389	0.0007
causa 4 : [65 − 70)	0.4941	0.1119	0.0000
causa 6 : [65 − 70)	-0.1032	0.1084	0.3410
causa 7 : [65 − 70)	2.5379	0.1159	0.0000
causa 8 : [65 − 70)	2.7510	0.1546	0.0000
causa 9 : [65 − 70)	3.4096	0.2414	0.0000
causa 10 : [65 − 70)	0.7785	0.1315	0.0000
causa 11 : [65 − 70)	2.6028	0.1549	0.0000
causa 12 : [65 − 70)	2.0052	0.1265	0.0000
causa 13 : [65 − 70)	3.4789	0.5111	0.0000
causa 14 : [65 − 70)	2.1704	0.2129	0.0000
causa 15 : [65 − 70)	-6.3158	0.1145	0.0000
causa 16 : [65 − 70)	0.7981	0.1094	0.0000
causa 17 : [65 − 70)	-0.0749	0.1058	0.4788
causa 2 : [70 − 75)	4.1692	0.1254	0.0000
causa 3 : [70 − 75)	-0.4549	0.1374	0.0009
causa 4 : [70 − 75)	0.5011	0.1115	0.0000
causa 6 : [70 − 75)	0.0967	0.1079	0.3704
causa 7 : [70 − 75)	2.6087	0.1156	0.0000
causa 8 : [70 − 75)	2.9261	0.1543	0.0000
causa 9 : [70 − 75)	3.4436	0.2411	0.0000
causa 10 : [70 − 75)	0.9880	0.1308	0.0000
causa 11 : [70 − 75)	2.7821	0.1546	0.0000
causa 12 : [70 − 75)	1.7734	0.1263	0.0000
causa 13 : [70 − 75)	3.6575	0.5108	0.0000
causa 14 : [70 − 75)	2.3900	0.2124	0.0000
causa 15 : [70 − 75)	-6.7226	0.1150	0.0000
causa 16 : [70 − 75)	0.6223	0.1091	0.0000

causa 17 : [70 − 75)	-0.2335	0.1055	0.0268
causa 2 : [75 − 80)	3.8733	0.1252	0.0000
causa 3 : [75 − 80)	-0.4801	0.1364	0.0004
causa 4 : [75 − 80)	0.4559	0.1111	0.0000
causa 6 : [75 − 80)	0.2682	0.1076	0.0126
causa 7 : [75 − 80)	2.6108	0.1154	0.0000
causa 8 : [75 − 80)	3.0013	0.1541	0.0000
causa 9 : [75 − 80)	3.3885	0.2410	0.0000
causa 10 : [75 − 80)	1.1919	0.1304	0.0000
causa 11 : [75 − 80)	2.8079	0.1544	0.0000
causa 12 : [75 − 80)	1.5062	0.1260	0.0000
causa 13 : [75 − 80)	3.8027	0.5106	0.0000
causa 14 : [75 − 80)	2.5643	0.2121	0.0000
causa 15 : [75 − 80)	-7.3169	0.1172	0.0000
causa 16 : [75 − 80)	0.4732	0.1088	0.0000
causa 17 : [75 − 80)	-0.3519	0.1052	0.0008
causa 2 : [80 − 85)	3.5733	0.1251	0.0000
causa 3 : [80 − 85)	-0.4531	0.1359	0.0009
causa 4 : [80 − 85)	0.4108	0.1110	0.0002
causa 6 : [80 − 85)	0.4083	0.1074	0.0001
causa 7 : [80 − 85)	2.6527	0.1153	0.0000
causa 8 : [80 − 85)	3.0416	0.1540	0.0000
causa 9 : [80 − 85)	3.3146	0.2409	0.0000
causa 10 : [80 − 85)	1.4463	0.1302	0.0000
causa 11 : [80 − 85)	2.7868	0.1543	0.0000
causa 12 : [80 − 85)	1.3306	0.1260	0.0000
causa 13 : [80 − 85)	3.9200	0.5105	0.0000
causa 14 : [80 − 85)	2.7868	0.2120	0.0000
causa 15 : [80 − 85)	-7.6719	0.1196	0.0000
causa 16 : [80 − 85)	0.4210	0.1088	0.0001
causa 17 : [80 − 85)	-0.3278	0.1051	0.0018
causa 2 : [85 − 90)	3.2886	0.1252	0.0000
causa 3 : [85 − 90)	-0.3544	0.1360	0.0092
causa 4 : [85 − 90)	0.4238	0.1111	0.0001
causa 6 : [85 − 90)	0.4462	0.1075	0.0000
causa 7 : [85 − 90)	2.7472	0.1153	0.0000
causa 8 : [85 − 90)	3.0635	0.1541	0.0000
causa 9 : [85 − 90)	3.3204	0.2410	0.0000
causa 10 : [85 − 90)	1.7426	0.1302	0.0000
causa 11 : [85 − 90)	2.7816	0.1544	0.0000
causa 12 : [85 − 90)	1.2835	0.1261	0.0000
causa 13 : [85 − 90)	4.0356	0.5105	0.0000
causa 14 : [85 − 90)	2.9642	0.2120	0.0000
causa 15 : [85 − 90)	-7.7988	0.1246	0.0000
causa 16 : [85 − 90)	0.4731	0.1088	0.0000
causa 17 : [85 − 90)	-0.1647	0.1051	0.1172
causa 2 : [90 − 95)	3.0192	0.1255	0.0000
causa 3 : [90 − 95)	-0.0654	0.1368	0.6324

causa 4 : [90 − 95)	0.5039	0.1116	0.0000
causa 6 : [90 − 95)	0.2666	0.1080	0.0135
causa 7 : [90 − 95)	2.8399	0.1156	0.0000
causa 8 : [90 − 95)	3.0632	0.1543	0.0000
causa 9 : [90 − 95)	3.3132	0.2412	0.0000
causa 10 : [90 − 95)	2.0829	0.1305	0.0000
causa 11 : [90 − 95)	2.7771	0.1547	0.0000
causa 12 : [90 − 95)	1.2751	0.1265	0.0000
causa 13 : [90 − 95)	4.1470	0.5107	0.0000
causa 14 : [90 − 95)	3.1338	0.2122	0.0000
causa 15 : [90 − 95)	-8.0134	0.1470	0.0000
causa 16 : [90 − 95)	0.5683	0.1092	0.0000
causa 17 : [90 − 95)	0.1134	0.1055	0.2822
causa 2 : [95 e oltre)	2.7312	0.1269	0.0000
causa 3 : [95 e oltre)	0.1419	0.1407	0.3134
causa 4 : [95 e oltre)	0.6703	0.1134	0.0000
causa 6 : [95 e oltre)	0.0115	0.1101	0.9165
causa 7 : [95 e oltre)	2.9366	0.1170	0.0000
causa 8 : [95 e oltre)	2.9558	0.1556	0.0000
causa 9 : [95 e oltre)	3.2692	0.2424	0.0000
causa 10 : [95 e oltre)	2.4168	0.1319	0.0000
causa 11 : [95 e oltre)	2.8820	0.1559	0.0000
causa 12 : [95 e oltre)	1.2693	0.1284	0.0000
causa 13 : [95 e oltre)	4.3040	0.5114	0.0000
causa 14 : [95 e oltre)	3.2768	0.2133	0.0000
causa 15 : [95 e oltre)	-8.0483	0.2100	0.0000
causa 16 : [95 e oltre)	0.7053	0.1110	0.0000
causa 17 : [95 e oltre)	0.5605	0.1070	0.0000
$log(\theta)$	3.5641	0.0282	0.0000
intercetta	-22.4130	1.0548	0.0000

Tabella A.1: *Output* del modello ZINB relativo agli uomini senza la causa 5 (le celle colorate individuano i *p-value* superiori a 0.01 i cui parametri non sono quindi significativamente diversi da 0). La prima parte (prima della riga) è riferita alla distribuzione Binomiale Negativa mentra la seconda alla Binomiale.

A.2 Il GLM con risposta Binomiale Negativa per le donne

	Donne		
	Stima	*Std Error*	*p-value*
causa 1	-9.0543	0.0926	0.0000
causa 2	-10.0125	0.1353	0.0000
causa 3	-10.4210	0.1791	0.0000
causa 4	-9.2801	0.0935	0.0000
causa 5	-12.7994	1.0015	0.0000

causa 6	-8.6683	0.0721	0.0000
causa 7	-9.5259	0.1085	0.0000
causa 8	-10.2930	0.1955	0.0000
causa 9	-12.8512	0.5563	0.0000
causa 10	-10.3979	0.1699	0.0000
causa 11	-10.8333	0.2187	0.0000
causa 12	-9.9154	0.1415	0.0000
causa 13	-12.6083	0.5916	0.0000
causa 14	-11.8623	0.3840	0.0000
causa 15	-5.7699	0.0320	0.0000
causa 16	-8.9239	0.0831	0.0000
causa 17	-8.5764	0.0654	0.0000
$[1-5)$	-2.0526	0.1502	0.0000
$[5-10)$	-3.5031	0.2285	0.0000
$[10-15)$	-3.6629	0.2509	0.0000
$[15-20)$	-3.1637	0.1973	0.0000
$[20-25)$	-3.0509	0.1867	0.0000
$[25-30)$	-2.4227	0.1551	0.0000
$[30-35)$	-1.4683	0.1232	0.0000
$[35-40)$	-0.8613	0.1120	0.0000
$[40-45)$	-0.9261	0.1110	0.0000
$[45-50)$	-1.0276	0.1111	0.0000
$[50-55)$	-1.0452	0.1107	0.0000
$[55-60)$	-0.6081	0.1086	0.0000
$[60-65)$	-0.1081	0.1058	0.3071
$[65-70)$	0.3689	0.1024	0.0003
$[70-75)$	1.0327	0.0999	0.0000
$[75-80)$	1.6611	0.0986	0.0000
$[80-85)$	2.2297	0.0983	0.0000
$[85-90)$	2.9009	0.0980	0.0000
$[90-95)$	3.4312	0.0981	0.0000
[95 e oltre)	3.9013	0.0996	0.0000
anno	-0.1031	0.0143	0.0000
anno^2	0.0010	0.0001	0.0000
causa 2 : $[1-5)$	1.9367	0.2160	0.0000
causa 3 : $[1-5)$	1.1061	0.2724	0.0000
causa 4 : $[1-5)$	-0.0131	0.2165	0.9516
causa 5 : $[1-5)$	1.0559	1.1505	0.3588
causa 6 : $[1-5)$	-0.0868	0.1949	0.6559
causa 7 : $[1-5)$	0.1117	0.2316	0.6296
causa 8 : $[1-5)$	-1.3583	0.4902	0.0056
causa 9 : $[1-5)$	0.2906	1.0085	0.7733
causa 10 : $[1-5)$	0.2477	0.2963	0.4032
causa 11 : $[1-5)$	0.5860	0.3389	0.0839
causa 12 : $[1-5)$	-0.3211	0.3046	0.2918
causa 13 : $[1-5)$	0.7886	0.7804	0.3122
causa 14 : $[1-5)$	-0.2146	0.6729	0.7498
causa 15 : $[1-5)$	-2.7038	0.1774	0.0000

causa 16 : [1 − 5)	1.5272	0.1814	0.0000
causa 17 : [1 − 5)	-0.1090	0.1905	0.5673
causa 2 : [5 − 10)	3.1641	0.2759	0.0000
causa 3 : [5 − 10)	0.8901	0.3911	0.0228
causa 4 : [5 − 10)	0.6512	0.2994	0.0296
causa 5 : [5 − 10)	1.3890	1.2259	0.2572
causa 6 : [5 − 10)	0.6215	0.2716	0.0221
causa 7 : [5 − 10)	0.5412	0.3307	0.1017
causa 8 : [5 − 10)	0.1924	0.4494	0.6685
causa 9 : [5 − 10)	1.3687	0.9994	0.1708
causa 10 : [5 − 10)	-0.4934	0.5337	0.3553
causa 11 : [5 − 10)	0.6016	0.4906	0.2200
causa 12 : [5 − 10)	-0.1012	0.4392	0.8178
causa 13 : [5 − 10)	1.8927	0.7875	0.0162
causa 14 : [5 − 10)	0.2929	0.8687	0.7360
causa 15 : [5 − 10)	-2.4838	0.2747	0.0000
causa 16 : [5 − 10)	2.2214	0.2543	0.0000
causa 17 : [5 − 10)	0.0650	0.2841	0.8190
causa 2 : [10 − 15)	3.1860	0.2954	0.0000
causa 3 : [10 − 15)	1.2138	0.3883	0.0018
causa 4 : [10 − 15)	1.1283	0.3057	0.0002
causa 5 : [10 − 15)	2.6612	1.1018	0.0157
causa 6 : [10 − 15)	0.8101	0.2896	0.0051
causa 7 : [10 − 15)	1.1604	0.3215	0.0003
causa 8 : [10 − 15)	0.7944	0.4254	0.0619
causa 9 : [10 − 15)	2.3664	0.7851	0.0026
causa 10 : [10 − 15)	0.5003	0.4705	0.2876
causa 11 : [10 − 15)	1.4193	0.4279	0.0009
causa 12 : [10 − 15)	1.0703	0.3730	0.0041
causa 13 : [10 − 15)	2.9453	0.7102	0.0000
causa 14 : [10 − 15)	-0.5597	1.1554	0.6281
causa 15 : [10 − 15)	-2.3273	0.2907	0.0000
causa 16 : [10 − 15)	2.6165	0.2723	0.0000
causa 17 : [10 − 15)	0.3203	0.2978	0.2820
causa 2 : [15 − 20)	3.0388	0.2484	0.0000
causa 3 : [15 − 20)	0.7420	0.3541	0.0361
causa 4 : [15 − 20)	0.7765	0.2562	0.0024
causa 5 : [15 − 20)	2.4334	1.0616	0.0219
causa 6 : [15 − 20)	0.8580	0.2312	0.0002
causa 7 : [15 − 20)	1.2803	0.2550	0.0000
causa 8 : [15 − 20)	0.6500	0.3524	0.0651
causa 9 : [15 − 20)	2.3538	0.6939	0.0007
causa 10 : [15 − 20)	0.3888	0.3724	0.2964
causa 11 : [15 − 20)	1.5729	0.3567	0.0000
causa 12 : [15 − 20)	0.3393	0.3365	0.3133
causa 13 : [15 − 20)	2.7337	0.6804	0.0001
causa 14 : [15 − 20)	1.4830	0.5629	0.0084
causa 15 : [15 − 20)	-2.3804	0.2327	0.0000

causa 16 : [15 − 20)	3.4476	0.2181	0.0000
causa 17 : [15 − 20)	0.6619	0.2289	0.0038
causa 2 : [20 − 25)	3.1783	0.2384	0.0000
causa 3 : [20 − 25)	0.8614	0.3409	0.0115
causa 4 : [20 − 25)	0.8650	0.2409	0.0003
causa 5 : [20 − 25)	3.5142	1.0338	0.0007
causa 6 : [20 − 25)	0.8430	0.2198	0.0001
causa 7 : [20 − 25)	1.3237	0.2419	0.0000
causa 8 : [20 − 25)	1.3432	0.3174	0.0000
causa 9 : [20 − 25)	2.8648	0.6451	0.0000
causa 10 : [20 − 25)	0.6280	0.3486	0.0716
causa 11 : [20 − 25)	1.8153	0.3333	0.0000
causa 12 : [20 − 25)	0.2460	0.3114	0.4295
causa 13 : [20 − 25)	2.4757	0.6795	0.0003
causa 14 : [20 − 25)	2.4724	0.4696	0.0000
causa 15 : [20 − 25)	-2.7177	0.2268	0.0000
causa 16 : [20 − 25)	3.5064	0.2081	0.0000
causa 17 : [20 − 25)	1.4172	0.2095	0.0000
causa 2 : [25 − 30)	2.9585	0.2118	0.0000
causa 3 : [25 − 30)	0.0720	0.3227	0.8235
causa 4 : [25 − 30)	0.0421	0.2203	0.8486
causa 5 : [25 − 30)	3.3194	1.0233	0.0012
causa 6 : [25 − 30)	0.2124	0.1932	0.2716
causa 7 : [25 − 30)	1.0069	0.2100	0.0000
causa 8 : [25 − 30)	1.3054	0.2783	0.0000
causa 9 : [25 − 30)	2.6803	0.6169	0.0000
causa 10 : [25 − 30)	0.1247	0.3137	0.6911
causa 11 : [25 − 30)	1.2271	0.3108	0.0001
causa 12 : [25 − 30)	0.5297	0.2536	0.0367
causa 13 : [25 − 30)	2.4390	0.6506	0.0002
causa 14 : [25 − 30)	2.3647	0.4367	0.0000
causa 15 : [25 − 30)	-3.4399	0.2031	0.0000
causa 16 : [25 − 30)	2.7608	0.1804	0.0000
causa 17 : [25 − 30)	0.9108	0.1799	0.0000
causa 2 : [30 − 35)	2.7001	0.1873	0.0000
causa 3 : [30 − 35)	-0.6356	0.2915	0.0292
causa 4 : [30 − 35)	-0.8403	0.1945	0.0000
causa 5 : [30 − 35)	3.0453	1.0143	0.0027
causa 6 : [30 − 35)	-0.5430	0.1632	0.0009
causa 7 : [30 − 35)	0.4583	0.1804	0.0111
causa 8 : [30 − 35)	0.9171	0.2479	0.0002
causa 9 : [30 − 35)	2.5186	0.5885	0.0000
causa 10 : [30 − 35)	-0.3528	0.2710	0.1930
causa 11 : [30 − 35)	0.4852	0.2865	0.0903
causa 12 : [30 − 35)	0.4711	0.2093	0.0244
causa 13 : [30 − 35)	1.5262	0.6339	0.0161
causa 14 : [30 − 35)	2.0497	0.4166	0.0000
causa 15 : [30 − 35)	-4.2357	0.1707	0.0000

causa 16 : [30 − 35)	1.9239	0.1534	0.0000
causa 17 : [30 − 35)	0.1832	0.1505	0.2236
causa 2 : [35 − 40)	2.8864	0.1789	0.0000
causa 3 : [35 − 40)	-1.1009	0.2757	0.0001
causa 4 : [35 − 40)	-1.1331	0.1766	0.0000
causa 5 : [35 − 40)	3.1966	1.0105	0.0016
causa 6 : [35 − 40)	-0.5715	0.1472	0.0001
causa 7 : [35 − 40)	0.4140	0.1667	0.0130
causa 8 : [35 − 40)	0.8247	0.2358	0.0005
causa 9 : [35 − 40)	2.3233	0.5793	0.0001
causa 10 : [35 − 40)	-0.7541	0.2492	0.0025
causa 11 : [35 − 40)	0.4617	0.2679	0.0848
causa 12 : [35 − 40)	0.8677	0.1903	0.0000
causa 13 : [35 − 40)	1.3910	0.6216	0.0252
causa 14 : [35 − 40)	1.4535	0.4133	0.0004
causa 15 : [35 − 40)	-4.9933	0.1670	0.0000
causa 16 : [35 − 40)	1.4731	0.1441	0.0000
causa 17 : [35 − 40)	-0.0925	0.1388	0.5049
causa 2 : [40 − 45)	3.6079	0.1778	0.0000
causa 3 : [40 − 45)	-0.4536	0.2513	0.0711
causa 4 : [40 − 45)	-0.4977	0.1651	0.0026
causa 5 : [40 − 45)	3.7109	1.0095	0.0002
causa 6 : [40 − 45)	-0.3095	0.1444	0.0321
causa 7 : [40 − 45)	1.0741	0.1624	0.0000
causa 8 : [40 − 45)	1.4033	0.2318	0.0000
causa 9 : [40 − 45)	2.9211	0.5744	0.0000
causa 10 : [40 − 45)	-0.0745	0.2311	0.7471
causa 11 : [40 − 45)	0.8655	0.2608	0.0009
causa 12 : [40 − 45)	1.8026	0.1855	0.0000
causa 13 : [40 − 45)	1.3610	0.6202	0.0282
causa 14 : [40 − 45)	1.4526	0.4145	0.0005
causa 15 : [40 − 45)	-4.9611	0.1649	0.0000
causa 16 : [40 − 45)	1.7299	0.1430	0.0000
causa 17 : [40 − 45)	0.1604	0.1367	0.2407
causa 2 : [45 − 50)	4.2550	0.1776	0.0000
causa 3 : [45 − 50)	-0.0506	0.2412	0.8339
causa 4 : [45 − 50)	0.2253	0.1576	0.1529
causa 5 : [45 − 50)	4.2148	1.0090	0.0000
causa 6 : [45 − 50)	0.1018	0.1420	0.4732
causa 7 : [45 − 50)	1.7013	0.1606	0.0000
causa 8 : [45 − 50)	1.8805	0.2302	0.0000
causa 9 : [45 − 50)	3.5125	0.5720	0.0000
causa 10 : [45 − 50)	0.0952	0.2270	0.6750
causa 11 : [45 − 50)	1.6055	0.2543	0.0000
causa 12 : [45 − 50)	2.3211	0.1842	0.0000
causa 13 : [45 − 50)	2.2381	0.6122	0.0003
causa 14 : [45 − 50)	1.5408	0.4143	0.0002
causa 15 : [45 − 50)	-4.5628	0.1545	0.0000

causa 16 : [45 − 50)	1.9510	0.1428	0.0000
causa 17 : [45 − 50)	0.5159	0.1357	0.0001
causa 2 : [50 − 55)	4.6315	0.1772	0.0000
causa 3 : [50 − 55)	0.1854	0.2361	0.4323
causa 4 : [50 − 55)	0.6901	0.1538	0.0000
causa 5 : [50 − 55)	4.4488	1.0088	0.0000
causa 6 : [50 − 55)	0.3305	0.1401	0.0183
causa 7 : [50 − 55)	2.1052	0.1594	0.0000
causa 8 : [50 − 55)	2.1726	0.2291	0.0000
causa 9 : [50 − 55)	3.5499	0.5715	0.0000
causa 10 : [50 − 55)	0.4676	0.2211	0.0345
causa 11 : [50 − 55)	1.8998	0.2519	0.0000
causa 12 : [50 − 55)	2.6893	0.1833	0.0000
causa 13 : [50 − 55)	2.5201	0.6097	0.0000
causa 14 : [50 − 55)	1.8926	0.4102	0.0000
causa 15 : [50 − 55)	-4.1127	0.1449	0.0000
causa 16 : [50 − 55)	1.9904	0.1425	0.0000
causa 17 : [50 − 55)	0.6800	0.1347	0.0000
causa 2 : [55 − 60)	4.5309	0.1759	0.0000
causa 3 : [55 − 60)	-0.0778	0.2347	0.7404
causa 4 : [55 − 60)	0.7739	0.1510	0.0000
causa 5 : [55 − 60)	4.0900	1.0087	0.0001
causa 6 : [55 − 60)	0.2622	0.1379	0.0573
causa 7 : [55 − 60)	2.0992	0.1577	0.0000
causa 8 : [55 − 60)	2.0277	0.2280	0.0000
causa 9 : [55 − 60)	3.4533	0.5708	0.0000
causa 10 : [55 − 60)	0.3521	0.2180	0.1062
causa 11 : [55 − 60)	1.7685	0.2504	0.0000
causa 12 : [55 − 60)	2.3682	0.1821	0.0000
causa 13 : [55 − 60)	2.5380	0.6078	0.0000
causa 14 : [55 − 60)	1.9562	0.4075	0.0000
causa 15 : [55 − 60)	-4.3033	0.1420	0.0000
causa 16 : [55 − 60)	1.5766	0.1412	0.0000
causa 17 : [55 − 60)	0.3308	0.1334	0.0132
causa 2 : [60 − 65)	4.3369	0.1742	0.0000
causa 3 : [60 − 65)	-0.1850	0.2296	0.4204
causa 4 : [60 − 65)	0.7688	0.1475	0.0000
causa 5 : [60 − 65)	3.5531	1.0085	0.0004
causa 6 : [60 − 65)	0.2956	0.1344	0.0278
causa 7 : [60 − 65)	2.2242	0.1552	0.0000
causa 8 : [60 − 65)	2.0332	0.2260	0.0000
causa 9 : [60 − 65)	3.4201	0.5694	0.0000
causa 10 : [60 − 65)	0.4964	0.2120	0.0192
causa 11 : [60 − 65)	1.8635	0.2477	0.0000
causa 12 : [60 − 65)	2.1014	0.1804	0.0000
causa 13 : [60 − 65)	2.7240	0.6051	0.0000
causa 14 : [60 − 65)	2.0415	0.4042	0.0000
causa 15 : [60 − 65)	-5.1057	0.1496	0.0000

causa 16 : [60 − 65)	1.2062	0.1392	0.0000
causa 17 : [60 − 65)	0.1698	0.1307	0.1937
causa 2 : [65 − 70)	4.1837	0.1721	0.0000
causa 3 : [65 − 70)	-0.1199	0.2207	0.5868
causa 4 : [65 − 70)	1.0288	0.1433	0.0000
causa 5 : [65 − 70)	3.4697	1.0078	0.0006
causa 6 : [65 − 70)	0.3747	0.1303	0.0040
causa 7 : [65 − 70)	2.4853	0.1523	0.0000
causa 8 : [65 − 70)	2.2991	0.2234	0.0000
causa 9 : [65 − 70)	3.5542	0.5677	0.0000
causa 10 : [65 − 70)	0.7125	0.2056	0.0005
causa 11 : [65 − 70)	2.1230	0.2447	0.0000
causa 12 : [65 − 70)	1.8848	0.1780	0.0000
causa 13 : [65 − 70)	2.8711	0.6029	0.0000
causa 14 : [65 − 70)	2.3645	0.4007	0.0000
causa 15 : [65 − 70)	-5.6795	0.1524	0.0000
causa 16 : [65 − 70)	0.9290	0.1362	0.0000
causa 17 : [65 − 70)	0.0939	0.1268	0.4591
causa 2 : [70 − 75)	3.8684	0.1706	0.0000
causa 3 : [70 − 75)	-0.1789	0.2141	0.4033
causa 4 : [70 − 75)	1.0272	0.1406	0.0000
causa 5 : [70 − 75)	3.4937	1.0073	0.0005
causa 6 : [70 − 75)	0.4110	0.1273	0.0012
causa 7 : [70 − 75)	2.5835	0.1504	0.0000
causa 8 : [70 − 75)	2.4389	0.2218	0.0000
causa 9 : [70 − 75)	3.7002	0.5665	0.0000
causa 10 : [70 − 75)	0.9224	0.2014	0.0000
causa 11 : [70 − 75)	2.1539	0.2429	0.0000
causa 12 : [70 − 75)	1.6136	0.1763	0.0000
causa 13 : [70 − 75)	3.0126	0.6015	0.0000
causa 14 : [70 − 75)	2.3110	0.3990	0.0000
causa 15 : [70 − 75)	-5.9443	0.1439	0.0000
causa 16 : [70 − 75)	0.6741	0.1339	0.0000
causa 17 : [70 − 75)	0.0474	0.1239	0.7018
causa 2 : [75 − 80)	3.5564	0.1698	0.0000
causa 3 : [75 − 80)	-0.1850	0.2105	0.3795
causa 4 : [75 − 80)	1.0464	0.1393	0.0000
causa 5 : [75 − 80)	3.8661	1.0069	0.0001
causa 6 : [75 − 80)	0.6100	0.1258	0.0000
causa 7 : [75 − 80)	2.7712	0.1494	0.0000
causa 8 : [75 − 80)	2.7017	0.2210	0.0000
causa 9 : [75 − 80)	3.8284	0.5660	0.0000
causa 10 : [75 − 80)	1.2830	0.1994	0.0000
causa 11 : [75 − 80)	2.1694	0.2420	0.0000
causa 12 : [75 − 80)	1.5230	0.1752	0.0000
causa 13 : [75 − 80)	3.2330	0.6008	0.0000
causa 14 : [75 − 80)	2.4509	0.3981	0.0000
causa 15 : [75 − 80)	-6.4255	0.1409	0.0000

causa 16 : [75 − 80)	0.6232	0.1326	0.0000
causa 17 : [75 − 80)	0.1500	0.1223	0.2197
causa 2 : [80 − 85)	3.3435	0.1697	0.0000
causa 3 : [80 − 85)	-0.0410	0.2094	0.8447
causa 4 : [80 − 85)	1.1321	0.1390	0.0000
causa 5 : [80 − 85)	4.2879	1.0068	0.0000
causa 6 : [80 − 85)	0.7318	0.1254	0.0000
causa 7 : [80 − 85)	3.0094	0.1492	0.0000
causa 8 : [80 − 85)	2.8948	0.2208	0.0000
causa 9 : [80 − 85)	4.0505	0.5658	0.0000
causa 10 : [80 − 85)	1.7395	0.1988	0.0000
causa 11 : [80 − 85)	2.2354	0.2418	0.0000
causa 12 : [80 − 85)	1.5731	0.1750	0.0000
causa 13 : [80 − 85)	3.4438	0.6006	0.0000
causa 14 : [80 − 85)	2.7046	0.3978	0.0000
causa 15 : [80 − 85)	-6.5559	0.1430	0.0000
causa 16 : [80 − 85)	0.7812	0.1322	0.0000
causa 17 : [80 − 85)	0.4175	0.1218	0.0006
causa 2 : [85 − 90)	3.0027	0.1695	0.0000
causa 3 : [85 − 90)	0.0517	0.2084	0.8041
causa 4 : [85 − 90)	1.1738	0.1386	0.0000
causa 5 : [85 − 90)	4.6699	1.0067	0.0000
causa 6 : [85 − 90)	0.6341	0.1251	0.0000
causa 7 : [85 − 90)	3.1690	0.1489	0.0000
causa 8 : [85 − 90)	2.9719	0.2206	0.0000
causa 9 : [85 − 90)	4.2473	0.5657	0.0000
causa 10 : [85 − 90)	2.1079	0.1984	0.0000
causa 11 : [85 − 90)	2.2786	0.2416	0.0000
causa 12 : [85 − 90)	1.6026	0.1747	0.0000
causa 13 : [85 − 90)	3.5939	0.6005	0.0000
causa 14 : [85 − 90)	2.8027	0.3976	0.0000
causa 15 : [85 − 90)	-7.0668	0.1500	0.0000
causa 16 : [85 − 90)	0.8855	0.1318	0.0000
causa 17 : [85 − 90)	0.7382	0.1214	0.0000
causa 2 : [90 − 95)	2.7381	0.1696	0.0000
causa 3 : [90 − 95)	0.1798	0.2082	0.3879
causa 4 : [90 − 95)	1.2902	0.1387	0.0000
causa 5 : [90 − 95)	4.9692	1.0067	0.0000
causa 6 : [90 − 95)	0.4630	0.1253	0.0002
causa 7 : [90 − 95)	3.3150	0.1490	0.0000
causa 8 : [90 − 95)	2.9893	0.2207	0.0000
causa 9 : [90 − 95)	4.3833	0.5657	0.0000
causa 10 : [90 − 95)	2.4674	0.1984	0.0000
causa 11 : [90 − 95)	2.3389	0.2416	0.0000
causa 12 : [90 − 95)	1.7035	0.1748	0.0000
causa 13 : [90 − 95)	3.7655	0.6005	0.0000
causa 14 : [90 − 95)	2.8717	0.3976	0.0000
causa 15 : [90 − 95)	-7.5042	0.1736	0.0000

causa 16 : [90 − 95)	1.0469	0.1319	0.0000
causa 17 : [90 − 95)	1.1382	0.1214	0.0000
causa 2 : [95 e oltre)	2.4385	0.1707	0.0000
causa 3 : [95 e oltre)	0.3318	0.2106	0.1152
causa 4 : [95 e oltre)	1.4862	0.1400	0.0000
causa 5 : [95 e oltre)	5.1206	1.0069	0.0000
causa 6 : [95 e oltre)	0.1697	0.1271	0.1820
causa 7 : [95 e oltre)	3.4020	0.1500	0.0000
causa 8 : [95 e oltre)	2.8791	0.2215	0.0000
causa 9 : [95 e oltre)	4.4544	0.5661	0.0000
causa 10 : [95 e oltre)	2.7794	0.1992	0.0000
causa 11 : [95 e oltre)	2.4844	0.2425	0.0000
causa 12 : [95 e oltre)	1.7530	0.1760	0.0000
causa 13 : [95 e oltre)	3.8767	0.6009	0.0000
causa 14 : [95 e oltre)	2.9321	0.3983	0.0000
causa 15 : [95 e oltre)	-7.3515	0.2414	0.0000
causa 16 : [95 e oltre)	1.2199	0.1332	0.0000
causa 17 : [95 e oltre)	1.6565	0.1226	0.0000

Tabella A.2: *Output* parziale con gli effetti principali e le interazioni di primo grado tra le cause e la classe di età del GLM con risposta Binomiale Negativa per i dati relativi alle donne (le celle colorate individuano i *p-value* superiori a 0.01 i cui parametri non sono quindi significativamente diversi da 0).