

POLITECNICO DI MILANO

FACOLTA' DI INGEGNERIA DEI SISTEMI

Corso di Laurea Magistrale in INGEGNERIA MATEMATICA



**METODI GRAFICI ED INFERENZIALI
PER L'IDENTIFICAZIONE DI OUTLIERS:
IL CASO DEL PROCESSO DI CURA DI
PATOLOGIE CARDIOVASCOLARI**

Relatore:

Prof.ssa Anna Maria Paganoni

Tesi di Laurea di:

Emanuele Davide Giani

Matricola: 765923

Anno Accademico 2012-2013

Indice

Introduzione	v
1 Funnel plots	1
1.1 Brevi richiami statistici	2
1.2 Costruzione di un Funnel plot	4
1.2.1 Risposte continue	6
1.2.2 Proporzioni	7
1.2.3 <i>Standardized rates</i>	11
1.3 <i>Risk-adjusted data</i> : un esempio	14
1.3.1 Un'alternativa per il parametro di precisione	19
2 <i>Over-Dispersion</i>	23
2.1 Approccio moltiplicativo	25
2.1.1 Stima <i>Winsorised</i>	25
2.1.2 Selezione diretta	27
2.2 Approccio additivo	27
3 Utilizzo di Modelli Gerarchici	33
3.1 <i>Random effect model</i>	34
3.2 Tre possibili procedimenti	38
3.2.1 Approccio 1	40
3.2.2 Approccio 2	40
3.2.3 Approccio 3	41
3.3 L'influenza della correlazione	43
3.4 Un esempio	44
4 Dati del progetto <i>Scompenso</i> di Regione Lombardia	49
4.1 Presentazione del data-set	50
4.2 Funnel Plots	53
4.2.1 Funnel Plot senza correzione per l'over-dispersion . . .	55
4.2.2 Correzioni per l'over-dispersion	56

4.3	Utilizzo del modello <i>random effect</i>	60
4.4	Un'osservazione prima di proseguire	64
5	GLM e Regressione Logistica	69
5.1	Principali elementi di un GLM	69
5.2	GLM per risposte binarie	71
5.2.1	Identity link: linear probability model	72
5.2.2	Canonical link: logistic regression model	72
5.2.3	GLM Binomiale per una risposta contro un predittore binario	73
5.2.4	Inverse CDF link function: modello probit	73
5.3	Media e Varianza per la componente casuale	74
5.4	Inferenza e comparazione tra modelli	76
5.4.1	Deviance	76
5.4.2	Comparazione di modelli tramite la deviance	77
5.4.3	Comparazione di modelli tramite l'AIC	78
5.4.4	Residui per i GLM	79
5.5	La Regressione Logistica	81
5.5.1	Interpretazione dei coefficienti e forma della curva	82
5.5.2	Inferenza	82
5.5.3	Regressione logistica con un predittore categoriale	85
5.6	Regressione logistica multipla	86
5.7	Analisi dei dati del progetto <i>Scompenso</i>	86
6	Generalized Linear Mixed Models	97
6.1	Elementi di un GLMM	98
6.2	Logistic-Normal Model	99
6.2.1	Modello condizionato e modello marginale	100
6.2.2	Cenni sull'inferenza per i GLMM	101
6.3	Un modello GLMM per i dati HCD	101
7	Funnel Plot e GLMM a confronto	107
7.1	Indici di performance, di precisione e limiti di controllo	108
7.2	Funnel Plot senza correzioni	109
7.3	Funnel Plot corretti per l'over-dispersion	113
7.3.1	Correzione moltiplicativa	113
7.3.2	Correzione additiva	116
7.4	Funnel Plot e modello <i>random effect</i>	122
7.5	Confronto con il QQ-plot degli effetti casuali	132
	Conclusioni	137

<i>INDICE</i>	iii
Ringraziamenti	139
Bibliografia	141

Introduzione

LO SVILUPPO di metodi inferenziali per svolgere *outlier detection* riveste indubbiamente una grande utilità pratica: lo scopo è quello di permettere lo studio approfondito di un fenomeno, verificare le assunzioni di un modello statistico e indicare quali dati paiono in disaccordo rispetto al modello stesso. Poniamo il caso di uno studio riguardante le performances di numerose unità statistiche, ad esempio i licei di una città, i medici operanti in una clinica specializzata, gli ospedali di una regione italiana: l'interesse finale può essere quello di identificare unità particolarmente virtuose e di riconoscere invece le unità meno virtuose in un certo periodo in esame, con lo scopo ad esempio di premiare le eccellenze del settore o di intervenire per migliorare i provider meno efficienti.

In questo lavoro ci occupiamo per prima cosa di presentare una metodologia grafica adatta all'applicazione nella situazione appena descritta. Nei primi tre capitoli descriveremo nel dettaglio i *Funnel Plot*, indicando in maniera accurata come costruire questi grafici e spiegandone potenzialità e vantaggi rispetto ad altre metodologie grafiche.

Due capitoli saranno poi dedicati alla descrizione dei modelli lineari generalizzati (GLM) e generalizzati ad effetti misti (GLMM), al fine di confrontare ed integrare successivamente questi modelli con i Funnel Plot presentati nella prima parte di questo lavoro; ci interesseremo, attraverso lo studio di un caso pratico, di verificare se i metodi grafici introdotti e l'analisi dei modelli GLMM portino a conclusioni compatibili.

Questi confronti verranno svolti nell'ambito dello studio delle performances di alcuni ospedali operanti in Lombardia nell'ambito della cura di patologie di natura cardiovascolare: le procedure statistiche presentate in questo lavoro verranno dunque utilizzate per identificare le strutture caratterizzate da 'buone' performance e quelle caratterizzate invece da 'cattive' performance (in quale senso sarà opportunamente specificato). Infine sarà interessante rimarcare come il paradigma operativo presentato in questo lavoro si adatti a numerosissime altre situazioni pratiche.

Capitolo 1

Funnel plots

MOLTA attenzione viene oggi posta sulle metodologie per svolgere comparazioni sulle performance di strutture istituzionali basate su misure quantitative di outcome sensibili: risultati di esami scolastici, tasso di mortalità seguente ad operazioni chirurgiche, risultati di ricerca di diverse università. ‘Istituzioni’ può riferirsi a strutture ospedaliere o scolastiche, a team di chirurghi o a singoli chirurghi.

Queste comparazioni portano generalmente alla produzione di ‘classifiche’ dove le unità in analisi sono ordinate seguendo un indicatore di bontà della performance (spesso corredati da intervalli di confidenza). Questi grafici sono spesso chiamati **grafici caterpillar**. Questo modo di presentare i risultati non è esente da critiche, in quanto porta alla costruzione di un ordinamento tra le unità analizzate; ci si focalizza di conseguenza su una quantità (il *rank order*) tra le più difficili da stimare, ma soprattutto da costruire in maniera sensata e consistente.

Vogliamo quindi presentare uno strumento alternativo per il confronto di performance tra diverse unità: il *Funnel Plot*¹, nel quale un indicatore di performance osservato è plottato in funzione di una misura della sua precisione. L’aggiunta di opportuni limiti di controllo genera conseguentemente la tipica struttura ad imbuto del grafico, che permette ad un primo sguardo di suddividere le unità analizzate tra unità sotto controllo, e unità che presentino performance inusuali o addirittura estreme.

Come fortemente rimarcato in [13] e [14], un approccio di questo tipo ha indiscutibilmente numerosi vantaggi, primo tra i quali la facile interpretazione: lo sguardo risulta infatti guidato verso quei punti che giacciono all’esterno dei limiti di controllo. Tali grafici permettono anche uno studio informale del legame tra prestazione e precisione, e rendono poi possibile la visualizzazione

¹Funnel=imbuto.

dei progressi nel caso di misure ripetute nel tempo. Evitano poi di creare una classificazione fuorviante tra le strutture studiate, oltre al fatto che sono facilmente costruibili con i software più diffusi.

1.1 Brevi richiami statistici

Premettiamo, prima di proseguire, tre brevi richiami di calcolo delle probabilità utilizzati nel seguente paragrafo. Cominciamo con la seguente

Proposizione 1.1 (Formula di propagazione dell'errore). *Sia X una variabile aleatoria e $Y = h(X)$ una sua generica trasformazione non lineare. Se h è derivabile allora*

$$\mathbb{V}(Y) \approx \left(\frac{dh}{dX} \right)_{\mu_X}^2 \sigma_X^2$$

dove μ_X e σ_X^2 rappresentano valore atteso e varianza di X ; la derivata di h è valutata in μ_X .

Dimostrazione. Supponendo che la funzione h sia derivabile, possiamo scrivere lo sviluppo di Taylor al primo ordine di Y nel punto μ_X :

$$Y \approx h(\mu_X) + \frac{dh}{dX}(X - \mu_X)$$

dove dh/dX è valutata in μ_X . Applicando l'operatore di varianza si ha:

$$\mathbb{V}(Y) \approx \mathbb{V}(h(\mu_X)) + \mathbb{V}\left(\frac{dh}{dX}(X - \mu_X)\right) = \left(\frac{dh}{dX}\right)^2 \sigma_X^2.$$

Vedi [8], pg 110-111. □

In seguito verrà utilizzata la distribuzione di Poisson e conseguentemente le seguenti due proprietà:

Proposizione 1.2 (Somma di Poisson indipendenti). *Siano X e Y due variabili aleatorie di Poisson indipendenti di parametri λ_1 e λ_2 rispettivamente. Allora $X + Y$ è distribuita come una Poisson di parametro $\lambda_1 + \lambda_2$.*

Dimostrazione. Poiché l'evento $\{X + Y = n\}$ può essere scritto come unione degli eventi disgiunti $\{X = k, Y = n - k\}$, $0 \leq k \leq n$, abbiamo

$$\begin{aligned}
 \mathbb{P}(X + Y = n) &= \sum_{k=0}^n \mathbb{P}(X = k, Y = n - k) & (1.1) \\
 &= \sum_{k=0}^n \mathbb{P}(X = k)\mathbb{P}(Y = n - k) \\
 &= \sum_{k=0}^n e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\
 &= e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^n \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
 &= \frac{e^{-(\lambda_1 + \lambda_2)}}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\
 &= \frac{e^{-(\lambda_1 + \lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n
 \end{aligned}$$

Quindi, $X + Y$ è distribuita come una Poisson di parametro $\lambda_1 + \lambda_2$. Vedi [10], esempio 3e pg 277. \square

Proposizione 1.3. *Siano X e Y due variabili aleatorie di Poisson indipendenti di parametri λ_1 e λ_2 rispettivamente. Allora la distribuzione di X condizionata a $X + Y = n$ è una Binomiale di parametri $\lambda_1/(\lambda_1 + \lambda_2)$ e n .*

Dimostrazione. Calcoliamo la legge condizionata nel seguente modo:

$$\begin{aligned}
 \mathbb{P}(X = k | X + Y = n) &= \frac{\mathbb{P}(X = k, X + Y = n)}{\mathbb{P}(X + Y = n)} & (1.2) \\
 &= \frac{\mathbb{P}(X = k, Y = n - k)}{\mathbb{P}(X + Y = n)} \\
 &= \frac{\mathbb{P}(X = k)\mathbb{P}(Y = n - k)}{\mathbb{P}(X + Y = n)}
 \end{aligned}$$

L'ultima uguaglianza è data dall'indipendenza di X e Y . Grazie alla **Proposizione 1.2** possiamo scrivere

$$\begin{aligned}
 \mathbb{P}(X = k | X + Y = n) &= \frac{e^{-\lambda_1} \lambda_1^k}{k!} \frac{e^{-\lambda_2} \lambda_2^{n-k}}{(n-k)!} \left(\frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^n}{n!} \right)^{-1} & (1.3) \\
 &= \frac{n!}{(n-k)!k!} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} \\
 &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k}
 \end{aligned}$$

Quindi la legge di X condizionata a $X + Y = n$ è una Binomiale di parametri $\lambda_1/(\lambda_1 + \lambda_2)$ ed n . Vedi [10], esempio 4d pg 279. \square

1.2 Costruzione di un Funnel plot

Vediamo ora come costruire un grafico a imbuto per indagare e confrontare le performance di I differenti unità: come mostrato in [13], per costruire un Funnel Plot (di cui mostriamo un esempio in Figura 1.1) è necessario specificare 4 elementi basilari:

1. Un *indicatore* di performance Y .
2. Un *target* θ_0 , che specifica il valore desiderato per l'indicatore Y . Per le unità considerate sotto controllo si ha quindi $\mathbb{E}(Y|\theta_0) = \theta_0$.
3. Un parametro di *precisione* ρ che determina l'accuratezza con cui l'indicatore Y è stato misurato. Sotto l'ipotesi nulla (unità considerata sotto controllo) assumiamo

$$Y_i \stackrel{ind}{\sim} p(y|\theta_0, \rho).$$

Non costituiscono però un campione identicamente distribuito in quanto l'indice di precisione ρ varia tra le diverse unità sotto esame, permettendo quindi di esaltare le diverse performance. Nel contesto da noi considerato, la quantità ρ è costruita come proporzionale all'inverso della varianza dell'indicatore Y :

$$\rho = g(\theta_0)/\mathbb{V}(Y|\theta_0) \tag{1.4}$$

per qualche funzione g . La scelta di ρ è sotto alcuni versi arbitraria; è comunque preferibile che sia una quantità facilmente interpretabile².

4. Opportuni *limiti di controllo* inferiore e superiore $y_p(\theta_0, \rho)$, dove p rappresenta un *p-value*, ovvero una probabilità di eccedere il limite per un'unità sotto l'ipotesi nulla. Di conseguenza $y_p = F^{-1}(p)$, dove F^{-1} è l'inversa della cumulativa di $p(y|\theta_0, \rho)$. $p = 0.001, 0.999$ e $p = 0.025, 0.975$ sono valori ragionevoli e corrispondono circa ad intervalli di ampiezza tra 2 e 3 deviazioni standard. I limiti più stretti possono essere considerati come soglia di *warning*, mentre i limiti più larghi

²Successivamente vedremo che in alcuni casi emerge naturalmente che ρ sia assimilabile ad una dimensione campionaria (es. il volume di casi trattati in una struttura ospedaliera). In altri casi ρ può risultare dipendente da qualche parametro di disturbo che necessiti di essere stimato.

come soglia di *alarm*. Nel caso la distribuzione $p(y|\theta_0, \rho)$ sia discreta, potrebbe essere necessario ricorrere ad una interpolazione, sfruttando un qualche parametro di *continuous adjustment*.

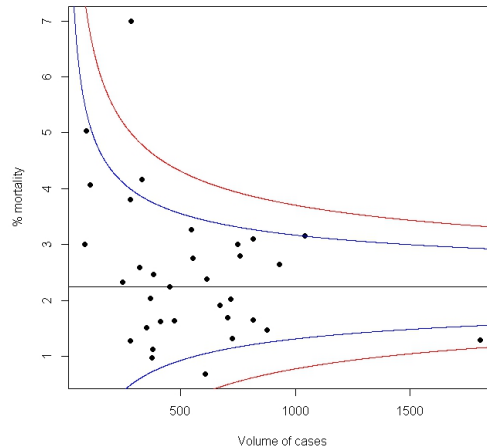


Figura 1.1: Tipico aspetto di un Funnel Plot, con la nuvola di punti rappresentanti le diverse unità e i limiti di controllo.

Data quindi una serie di I osservazioni y_i (indicatore di performance per l' i -esima unità) ad ognuna delle quali è associata una precisione ρ_i , un funnel plot consiste di un grafico delle y_i contro le precisioni ρ_i ; una linea orizzontale mette in evidenza il valore di target θ_0 . Sono aggiunti poi i limiti superiori e inferiori di controllo, plottati come funzione della precisione ρ . Come già detto, l'occhio viene quindi guidato da questi limiti di controllo verso quelle unità o strutture che cadono all'esterno di questi limiti, identificando quelle che presentano prestazioni particolarmente buone (o viceversa particolarmente cattive). Ad esempio in Figura 1.1 siamo immediatamente guidati ad osservare la struttura oltre il limite superiore, nell'angolo in alto a sinistra.

Un ulteriore elemento da specificare è la legge dell'indicatore Y per una unità sotto controllo. Chiaramente dipende dalla situazione in esame. In molti casi è possibile assumere una distribuzione (esatta o al più approssimata) normale:

$$Y|\theta_0, \rho \sim N(\theta_0, g(\theta_0)/\rho) \quad (1.5)$$

I limiti di controllo in questo caso particolare si esprimono quindi come segue:

$$y_p(\theta_0, \rho) = \theta_0 + z_p \sqrt{g(\theta_0)/\rho} \quad (1.6)$$

dove z_p è il punto percentuale di ordine p di una Normale standard. Vediamo di seguito come possono essere specificati questi elementi in alcuni casi specifici.

1.2.1 Risposte continue

Dati *Cross-sectional*

Supponiamo che per ogni unità sotto analisi (ad esempio una struttura ospedaliera) sia osservata una quantità di interesse su un campione di n casi. Indichiamo con y la media campionaria di ogni unità, affetta dall'errore standard σ .

1. L'indicatore è la media osservata y .
2. Il valore di target è scelto in modo che $\mathbb{V}(Y|\theta_0) = \sigma^2/n$.
3. Se si dispone di una stima comune della deviazione standard³ indicata con $\hat{\sigma}$, allora ρ è definito come la dimensione campionaria n . Nel caso non si disponesse di una stima della deviazione standard comune, si può prendere $\hat{\sigma}^2 = \sum n_i s_i^2 / I$, dove con s_i^2 si indica la varianza campionaria relativa a y_i , e si definisce ρ come l'*effective sample size*: per ogni struttura sotto osservazione abbiamo $n_i^{eff} = (\hat{\sigma}/s_i)^2$. In questo caso poi abbiamo $g(\theta_0) = \hat{\sigma}^2$.
4. Assumendo la distribuzione normale per Y , i limiti di controllo esatti si possono ottenere come $y_p = \theta_0 + z_p \sqrt{\hat{\sigma}^2/\rho}$.

Differenze in media

Supponiamo che per ogni unità sotto analisi siano disponibili due misurazioni y_1 e y_2 , rispettivamente con errore standard s_1 e s_2 (da non confondere con la deviazione standard campionaria indicata nella sezione precedente), relative a due periodi di osservazione successivi; il nostro interesse si focalizza su eventuali spostamenti della media tra i due periodi in esame.

³Assumiamo cioè che la risposta su ogni singolo caso osservato in una generica unità sotto esame sia affetta da deviazione standard σ .

1. L'indicatore è la differenza tra le medie $y = y_2 - y_1$.
2. Il target è specificato sulla differenza in media, e abbiamo $\mathbb{V}(Y|\theta_0) = s_1^2 + s_2^2$.
3. Una misura di precisione ρ può essere ottenuta nel seguente modo: supponendo di avere a disposizione una stima della deviazione standard comune si ha

$$\mathbb{V}(Y|\theta_0) = \hat{\sigma}^2/n_1 + \hat{\sigma}^2/n_2 = 2\hat{\sigma}^2 \frac{n_1 + n_2}{2n_1n_2} = g(\theta_0)/\rho.$$

Quindi $\rho = \frac{2n_1n_2}{n_1+n_2}$, che equivale a $\rho = n_1 + n_2$; nel caso che $n_1 = n_2 = N$, ρ è assimilabile al *sample size* del singolo periodo in esame. Al contempo abbiamo $g(\theta_0) = 2\hat{\sigma}^2$.

4. I limiti, assumendo come prima la distribuzione Normale, si ricavano immediatamente e valgono $y_p = \theta_0 + z_p \sqrt{2\hat{\sigma}^2/N}$.

1.2.2 Proporzioni

Dati *Cross-sectional*

Supponiamo che per ogni unità in esame siano osservati r eventi su un campione di dimensione n (ad esempio r pazienti morti su n pazienti curati in un reparto di una struttura ospedaliera). Di conseguenza si può ragionevolmente assumere per R una distribuzione Binomiale.

1. L'indicatore di interesse in questo caso è la proporzione osservata $y = r/n$.
2. La proporzione target θ_0 è tale che⁴ $\mathbb{V}(Y|\theta_0) = \theta_0(1 - \theta_0)/n$.
3. Una misura di facile interpretazione è quindi $\rho = n$, la dimensione campionaria. Dalla (1.1) abbiamo che $g(\theta_0) = \theta_0(1 - \theta_0)$.
4. Limiti di controllo esatti si possono ottenere dall'inversa della distribuzione Binomiale:

$$r_p = F^{-1}(p|\theta_0, n) \Leftrightarrow r_p = \inf \{ x \in \mathbb{N} \mid \mathbb{P}(R \leq x) = F(x|\theta_0, n) > p \} \quad (1.7)$$

⁴Essendo $R|\theta_0$ distribuita come una Binomiale di parametri n e θ_0 , abbiamo $\mathbb{V}(R|\theta_0) = n\theta_0(1 - \theta_0)$; da qui si ottiene immediatamente la varianza di Y .

dove F rappresenta la funzione di probabilità cumulativa di una Binomiale. Ciò significa che $F(r_p - 1|\theta_0, n) < p$; ma allora esiste un valore α tale che $\alpha F(r_p - 1|\theta_0, n) + (1 - \alpha)F(r_p|\theta_0, n) = p$, quindi

$$\alpha = \frac{F(r_p|\theta_0, n) - p}{F(r_p|\theta_0, n) - F(r_p - 1|\theta_0, n)}. \quad (1.8)$$

Il valore α ha quindi la funzione di un *continuous adjustment*, e risulta naturale prendere come limiti di controllo interpolati i seguenti valori:

$$y_p = (\alpha(r_p - 1) + (1 - \alpha)r_p)/n = (r_p - \alpha)/n. \quad (1.9)$$

È possibile ottenere anche i limiti normali approssimati direttamente dalla (1.3), e valgono $y_p = \theta_0 + z_p\sqrt{\theta_0(1 - \theta_0)/n}$.

Differenze tra le proporzioni

Supponiamo di disporre di due misurazioni per ogni unità: r_1, n_1 ed r_2, n_2 relativi a periodi di tempo successivi. Si è interessati a verificare un eventuale spostamento della proporzione dal valore π_1 al valore π_2 lungo l'arco di tempo di interesse. In questa situazione si hanno a disposizione tre possibili indicatori di performance: la differenza tra le proporzioni, il rapporto tra le proporzioni e l'*odds ratio*⁵. Vediamo nel dettaglio questi tre possibili approcci.

a. Differenza tra le proporzioni

1. L'indicatore di interesse in questo caso è la differenza tra le proporzioni osservate $y = r_2/n_2 - r_1/n_1$.
2. Il target θ_0 è posto sulla differenza $\pi_2 - \pi_1$. È immediato quindi verificare che, assumendo che R_1 sia Binomiale di parametri n_1 e π_1 e che R_2 sia Binomiale di parametri n_2 e π_2 ,

$$\mathbb{V}(Y|\pi_1, \pi_2) = \pi_2(1 - \pi_2)/n_2 + \pi_1(1 - \pi_1)/n_1.$$

⁵*Odds ratio (rapporto a favore)*. Per un evento A definiamo il rapporto a favore come:

$$\frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}.$$

Indica quanto è più probabile che si realizzi l'evento A rispetto al fatto che non si realizzi (vedi [10], pg 78).

Si può riparametrizzare l'espressione in termini della *proporzione media* $\pi_m = (\pi_1 + \pi_2)/2$; da questo si ha $\pi_2 = \pi_m + \theta_0/2$ e $\pi_1 = \pi_m - \theta_0/2$ e otteniamo

$$\mathbb{V}(Y|\theta_0, \pi_m) = \frac{(\pi_m + \theta_0/2)(1 - \pi_m - \theta_0/2)}{n_2} + \frac{(\pi_m - \theta_0/2)(1 - \pi_m + \theta_0/2)}{n_1}.$$

Il parametro *proporzione media* può essere stimato come

$$\hat{\pi}_m = (r_1 + r_2)/(n_1 + n_2).$$

3. Per ricavare un'espressione di facile interpretazione per la precisione, supponiamo che un'unità di interesse abbia sui due periodi in esame la medesima numerosità $n_1 = n_2 = N$ e proporzioni osservate pari entrambe alla proporzione media valutata sull'intero campione, stimata da $\hat{\pi}_m = \sum(r_{i1} + r_{i2})/(n_{i1} + n_{i2})$. Di conseguenza abbiamo $\mathbb{V}(Y|\theta_0, \hat{\pi}_m) = g(\theta_0)/N$, dove

$$g(\theta_0) = (\hat{\pi}_m + \theta_0/2)(1 - \hat{\pi}_m - \theta_0/2) + (\hat{\pi}_m - \theta_0/2)(1 - \hat{\pi}_m + \theta_0/2).$$

Come già in precedenza, possiamo interpretare il parametro di precisione ρ come la dimensione campionaria del singolo periodo.

4. In questo caso è poco agevole ricavare i limiti esatti; dalla (1.3) si possono ottenere subito i limiti di controllo normali approssimati.

b. Rapporto tra le proporzioni

1. L'indicatore è il rapporto $y = (r_2/n_2)/(r_1/n_1)$.
2. Il target θ_0 è posto sul rapporto tra le proporzioni π_2/π_1 , ma in questo caso è conveniente lavorare in scala logaritmica. Di conseguenza $\log(\theta_0)$ è un valore di target per $\log(Y)$. La **Proposizione 1.1** ci permette di scrivere $\mathbb{V}(Y|\pi_1, \pi_2) \approx (1 - \pi_2)/(n_2\pi_2) + (1 - \pi_1)/(n_1\pi_1)$, infatti:

$$\log Y = \log \frac{r_2}{n_2} - \log \frac{r_1}{n_1};$$

Quindi

$$\begin{aligned} \mathbb{V}\left(\log \frac{r_i}{n_i}\right) &\approx \left(\frac{n_i}{r_i} \frac{1}{n_i}\right)_{n_i\pi_i}^2 n_i\pi_i(1 - \pi_i) \\ &= \frac{1}{(n_i\pi_i)^2} n_i\pi_i(1 - \pi_i) \\ &= \frac{1 - \pi_i}{n_i\pi_i} \end{aligned}$$

che con i pari a 1 e 2 dà l'espressione approssimata scritta sopra. Riparametrizzando in termini della *proporzione media geometrica* $\pi_g = \sqrt{\pi_2\pi_1}$ abbiamo $\pi_2 = \pi_g\theta_0^{1/2}$ e $\pi_1 = \pi_g\theta_0^{-1/2}$, ottenendo infine

$$\mathbb{V}(\log Y|\theta_0, \pi_g) \approx \frac{\theta_0^{-1/2} - \pi_g}{n_2\pi_g} + \frac{\theta_0^{1/2} - \pi_g}{n_1\pi_g}.$$

Il valore π_g può essere stimato per ogni unità in analisi nel seguente modo: $\hat{\pi}_g = \sqrt{(r_1n_2)/(n_1r_2)}$.

3. Supponiamo che per un'unità si abbia $n_1 = n_2 = N$ e che le proporzioni osservate siano entrambe uguali alla media geometrica di tutto il campione, stimata da $\hat{\pi}_g = \sqrt{(\sum r_{i1} \sum r_{i2})/(\sum n_{i1} \sum n_{i2})}$. Allora possiamo scrivere $\mathbb{V}(\log Y|\theta_0, \hat{\pi}_g) = g(\theta_0)/N$, dove

$$g(\theta_0) = \frac{\theta_0^{-1/2} + \theta_0^{1/2} - 2\hat{\pi}_g}{\hat{\pi}_g}.$$

Ancora possiamo interpretare il parametro di precisione ρ come la dimensione campionaria del singolo periodo.

4. I limiti di controllo normali approssimati si ricavano ancora una volta dalla (1.3): essendo in scala logaritmica, andranno tracciati sul Funnel Plot adeguatamente riscalati.

c. Odds ratio per le proporzioni

1. In questo caso l'indicatore è l'*odds ratio* osservato $y = \frac{(r_2/(n_2-r_2))}{(r_1/(n_1-r_1))}$.
2. Il target θ_0 è posto sul *odds ratio* $(\pi_2/(1-\pi_2))/(\pi_1/(1-\pi_1))$; è però più conveniente lavorare in scala logaritmica, specificando quindi il target $\log \theta_0$ per la quantità $\log Y$. La **Proposizione 1.1** ci permette di scrivere

$$\mathbb{V}(\log Y) \approx \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)}.$$

Infatti:

$$\log Y = \log \frac{r_2}{n_2 - r_2} - \log \frac{r_1}{n_1 - r_1};$$

Quindi

$$\begin{aligned}
 \mathbb{V}\left(\log \frac{r_i}{n_i - r_i}\right) &\approx \left(\frac{n_i - r_i}{r_i} \frac{(n_i - r_i) + r_i}{(n_i - r_i)^2}\right)_{n_i \pi_i}^2 n_i \pi_i (1 - \pi_i) \\
 &= \left(\frac{n_i}{n_i \pi_i (n_i - n_i \pi_i)}\right)^2 n_i \pi_i (1 - \pi_i) \\
 &= \left(\frac{1}{n_i \pi_i (1 - \pi_i)}\right)^2 n_i \pi_i (1 - \pi_i) \\
 &= \frac{1}{n_i \pi_i (1 - \pi_i)}
 \end{aligned}$$

che per i pari a 1 e 2 dà l'espressione approssimata scritta sopra.

3. Supponendo che un'unità di interesse abbia $n_1 = n_2 = N$ e proporzioni osservate pari a $\hat{\pi}_m$, costruita come mostrato in precedenza abbiamo $\mathbb{V}(\log Y) \approx g(\theta_0)/N$, dove

$$g(\theta_0) = \frac{2}{\hat{\pi}_m(1 - \hat{\pi}_m)}.$$

Ecco che ρ è interpretabile ancora come la dimensione campionaria del singolo periodo in esame.

4. I limiti di controllo normali approssimati si ricavano ancora una volta dalla (1.3): essendo in scala logaritmica, andranno tracciati sul Funnel Plot adeguatamente riscaldati.

1.2.3 *Standardized rates*

Due tipi di dati riguardanti tassi standardizzati possono essere considerati: dati *indirettamente* standardizzati e dati *direttamente* standardizzati. Ci concentreremo solamente sul caso di dati *indirettamente* standardizzati i quali conducono alla costruzione di un *standardized event ratio* quale ad esempio un *standardized mortality ratio* (SMR). Questa procedura sarà ampiamente utilizzata nel seguito di questa trattazione.

Dati *Cross-sectional*

Supponiamo che in ogni unità siano osservati O eventi e che sia specificato in qualche modo (da informazioni a priori o da un modello statistico opportuno) un valore di *expectation* E .

1. L'indicatore è il valore osservato per l'SMR, $y = O/E$.
2. Il valore di target θ_0 per la quantità SMR è tale che una struttura sotto controllo avrà $O \sim Poisson(\theta_0 E)$. Di conseguenza $\mathbb{V}(Y|\theta_0) = \theta_0/E$.
3. Dall'espressione della varianza appena data, si trova che una misura facilmente interpretabile per la precisione è $\rho = E$. Quindi $g(\theta_0) = \theta_0$.
4. Limiti di controllo esatti si possono ottenere dall'inversa della distribuzione di Poisson con considerazioni simili a quelle utilizzate nel caso delle proporzioni:

$$o_p = F^{-1}(p|\theta_0 E) \Leftrightarrow o_p = \inf \{ x \in \mathbb{N} \mid \mathbb{P}(O \leq x) = F(x|\theta_0 E) > p \} \quad (1.10)$$

dove F rappresenta la funzione di probabilità cumulativa di una Poisson. Ciò significa che $F(o_p - 1|\theta_0 E) < p$; ma allora esiste un valore α tale che $\alpha F(o_p - 1|\theta_0 E) + (1 - \alpha)F(o_p|\theta_0 E) = p$, quindi

$$\alpha = \frac{F(o_p|\theta_0 E) - p}{F(o_p|\theta_0 E) - F(o_p - 1|\theta_0 E)}. \quad (1.11)$$

Il valore α ha ancora la funzione di un *continuous adjustment*, e risulta naturale prendere come limiti di controllo interpolati i seguenti valori:

$$y_p = (\alpha(o_p - 1) + (1 - \alpha)o_p)/E = (o_p - \alpha)/E. \quad (1.12)$$

È possibile ottenere anche i limiti normali approssimati direttamente dalla (1.3), e valgono $y_p = \theta_0 + z_p \sqrt{\theta_0/E}$.

Variazioni di *Standardized rates*

Supponiamo di avere due misure per ogni unità: O_1, E_1 in un primo periodo di tempo e O_2, E_2 in un seguente periodo di tempo. Il nostro obiettivo è valutare l'eventuale variazione nel tempo dell'SMR. Utilizziamo come indice il rapporto tra i tassi: un metodo esatto può essere ricavato nel caso particolare $E_1 = E_2 = E$, ovvero quando l'*expectation* è uguale per entrambi i periodi temporali in esame; ricaviamo anche un'approssimazione normale alternativa.

a. Risultato esatto ($E_1 = E_2 = E$)

1. L'indicatore è il rapporto tra gli SMR $y = (O_2/E_2)/(O_1/E_1) = O_2/O_1$.
2. Il target θ_0 è posto sul rapporto tra gli SMR, che indichiamo con λ_2/λ_1 .

3. La precisione ρ può essere definita in maniera naturale e molto comoda come il numero medio di eventi osservati: $\rho = (O_1 + O_2)/2$.
4. Per ricavare i limiti di controllo, partiamo coll'osservare che $O_1 \sim \text{Poisson}(\lambda_1 E)$ e $O_2 \sim \text{Poisson}(\theta_0 \lambda_1 E)$, con $O_1 \perp O_2$. Per la **Proposizione 1.3** si ha che la quantità $r = O_2 | O_1 + O_2 \sim B(\frac{\theta_0}{1+\theta_0}, O_1 + O_2)^6$. Ecco che sfruttando la (1.5) e la (1.6) è possibile ottenere i limiti $r_p(\theta_0/(1 + \theta_0), O_1 + O_2)$ per la proporzione $r = O_2/(O_1 + O_2)$; si può tornare alla scala originale applicando la trasformazione $y = r/(1 - r)$ per ottenere i limiti y_p .

b. Approssimazione normale

1. L'indicatore è lo stesso del caso precedente, ovvero il rapporto tra gli SMR osservati $y = (O_2/E_2)/(O_1/E_1)$.
2. Anche in questo caso il valore di target θ_0 è posto sul rapporto tra i tassi λ_2/λ_1 , ma risulta più conveniente lavorare in scala logaritmica. Utilizzando ancora la **Proposizione 1.1**, possiamo scrivere che $\mathbb{V}(\log Y | \lambda_1, \lambda_2) \approx 1/(E_2 \lambda_2) + 1/(E_1 \lambda_1)$. Infatti:

$$\log Y = \log \frac{O_2}{E_2} - \log \frac{O_1}{E_1};$$

Quindi

$$\begin{aligned} \mathbb{V}\left(\log \frac{O_i}{E_i}\right) &\approx \left(\frac{E_i}{O_i} \frac{1}{E_i}\right)_{E_i \lambda_i}^2 E_i \lambda_i \\ &= \left(\frac{1}{E_i \lambda_i}\right)_{E_i \lambda_i}^2 E_i \lambda_i \\ &= \frac{1}{E_i \lambda_i} \end{aligned} \tag{1.13}$$

che per i pari a 1 e 2 dà l'espressione approssimata scritta sopra. Se riparametrizziamo in termini della *media geometrica* $\lambda_g = \sqrt{\lambda_2 \lambda_1}$, tale per cui $\lambda_2 = \lambda_g \theta_0^{1/2}$ e $\lambda_1 = \lambda_g \theta_0^{-1/2}$, otteniamo

$$\mathbb{V}(\log Y | \theta_0, \lambda_g) = \frac{\theta_0^{-1/2}}{E_2 \lambda_g} + \frac{\theta_0^{1/2}}{E_1 \lambda_g}.$$

⁶Infatti per il primo parametro abbiamo:

$$\frac{\theta_0 \lambda_1}{\lambda_1 + \theta_0 \lambda_1} = \frac{\theta_0 \lambda_1}{\lambda_1(1 + \theta_0)} = \frac{\theta_0}{1 + \theta_0}.$$

Uno stimatore di massima verosimiglianza per λ_g può essere ottenuto sotto l'ipotesi nulla che $\lambda_2/\lambda_1 = \theta_0$: $\hat{\lambda}_g = (O_1 + O_2)/(\theta_0^{1/2}E_2 + \theta_0^{-1/2}E_1)$.

3. Per ricavare una misura interpretabile della precisione, supponiamo che per un'unità sotto esame si abbia che $E_1 = E_2 = E$ e che la media geometrica dell'SMR sia pari alla media geometrica stimata su tutto il campione, data da $\hat{\lambda}_g = (\sum O_{i1} + \sum O_{i2})/(\sum E_{i1} + \sum E_{i2})$. Per questa unità si avrebbe $\mathbb{V}(\log Y|\theta_0, \hat{\lambda}_g) = g(\theta_0)/E$, dove:

$$g(\theta_0) = (\theta_0^{-1/2} + \theta_0^{1/2})/\hat{\lambda}_g.$$

Ecco che ρ può essere quindi interpretato come l'*expectation* del singolo periodo.

4. I limiti di controllo normali approssimati si ricavano ancora una volta dalla (1.3): essendo in scala logaritmica, andranno tracciati sul Funnel plot adeguatamente riscaldati.

1.3 *Risk-adjusted data: un esempio*

In questa sezione illustriamo un esempio semplice ed istruttivo per mostrare le potenzialità dei Funnel Plot, introdotti in questo capitolo. In particolare introduciamo una procedura sfruttata nell'ambito del confronto di performance di diverse strutture ospedaliere al fine di *aggiustare secondo il rischio* i dati, ovvero di tenere conto nella valutazione della performance di altri parametri (quali il bacino di utenza, le patologie tipiche dei pazienti in un certo ospedale, etc.).

Utilizziamo i dati provenienti dal programma *CABG*⁷ dello stato di New York: sono resi disponibili i dati relativi alla mortalità di pazienti sottoposti a intervento di bypass coronarico con la procedura *CABG* in 33 ospedali operanti nello stato di New York in un periodo di 3 anni; in [17] sono pubblicati i dati relativi al triennio 1996-1998, mentre in [18] quelli relativi al triennio 1997-1999.

Gli outcomes di interesse mostrati in questi report sono:

1. Il numero di casi N (operazioni isolate) gestiti nel periodo in esame dal singolo ospedale.

⁷ *Coronary Artery Bypass Graft*: è una procedura chirurgica in cui una parte di vena o di arteria proveniente da un'altra parte del corpo del paziente è utilizzata per creare un percorso alternativo al fine di permettere al sangue di scorrere regolarmente verso il cuore, bypassando le arterie ostruite.

2. *Observed Mortality Rate* (OMR), equivalente a O/N , il numero di morti osservate nella struttura ospedaliera diviso i casi gestiti dalla stessa (presentato come percentuale nei report).
3. *Expacted Mortality Rate* (EMR), equivalente a E/N , consiste nella somma delle probabilità di morte per ogni paziente, previste attraverso un modello di regressione logistica, diviso per il numero di casi gestiti dalla struttura (presentato come percentuale nei report). Attraverso il modello di regressione è quindi possibile dare una previsione della probabilità di morte (e quindi delle morti attese in un certo ospedale) basata sulle caratteristiche del paziente (ad esempio: sesso, età, presenza di patologie quali ad esempio diabete o problemi renali, eventuali precedenti operazioni al cuore). Vedi [17] pg 7 e [18] pg 7 per i dettagli sulla regressione.
4. *Risk-adjusted Mortality Rate* (RAMR), equivalente a $\overline{OMR} \times (O/E) = \overline{OMR} \times (OMR/EMR)$, ovvero l'SMR che abbiamo già presentato precedentemente, moltiplicato per la media ponderata della mortalità osservata su tutti gli ospedali. Questa quantità è la migliore stima della mortalità che un ospedale dovrebbe presentare se gestisse un mix di pazienti rappresentativo delle caratteristiche di un paziente 'medio' presente nello stato di New York (presentato come percentuale nei report).

La procedura di aggiustamento per il rischio generalmente prende la forma di una standardizzazione indiretta. L'indicatore di performance considerato è l'RAMR. Osserviamo che

$$RAMR = \overline{OMR} \times \frac{O}{E} = \overline{OMR} \times Y$$

dove Y è l'indicatore che abbiamo già presentato nel paragrafo 1.2.3, e per il quale conosciamo già una misura corretta della precisione (l'*expectation* E) e l'espressione esatta o approssimata per la costruzione dei limiti di confidenza y_p . Per ottenere i limiti corretti sull'indicatore di interesse è sufficiente moltiplicare i limiti y_p per l'OMR medio, mentre per ottenere il numero di morti attese E è sufficiente moltiplicare N per EMR e dividere per 100.

In Tabella 1.1 e in Tabella 1.2 sono riportati i dati pubblicati rispettivamente in [17] e in [18].

Osserviamo i Funnel Plot ottenuti nei due casi (Figura 1.2 e Figura 1.3). In ognuna delle due figure, sono riportati a sinistra il Funnel plot con i limiti di tolleranza approssimati ricavati dalla formula (1.3), e a destra il Funnel plot con i limiti di tolleranza ricavati dalle formula (1.8) e (1.9). La prima cosa che

possiamo notare è che nel caso dei limiti di tolleranza esatti, il plot appare non liscio ma leggermente oscillante. Questo effetto è dovuto all'interpolazione introdotta dal parametro di *continuous adjustment* α . Altro aspetto che si può apprezzare è che i limiti sono molto simili, tanto simili che il limite esatto e il limite approssimato al 95% sono perfettamente sovrapposti, come mostrato in Figura 1.4 (linea grigia e linea blu) per entrambi i grafici.

La tipica forma a imbuto, come abbiamo già detto, permette di guidare il nostro sguardo direttamente verso le strutture che hanno presentato comportamenti notevoli. Nel periodo 1996-1998 osserviamo la presenza di una struttura (l'ospedale di Bellevue) che presenta un RAMR molto alto (oltre 12); ciò ci indurrebbe a pensare che questo ospedale è stato tra i meno virtuosi nel periodo osservato. Dato però il valore molto basso di morti attese per questa struttura, cioè una precisione molto bassa, la performance appare in realtà meno estrema di quanto non sembri. Strutture con RAMR più basso presentano in realtà comportamenti peggiori: per esempio l'ospedale di St. Vincents mostra un RAMR di 5.39; esso però cade abbondantemente oltre il limite superiore.

Ragionamenti analoghi possono essere fatti per il successivo periodo di tempo, dal 1997 al 1999. In particolare si vede anche in questo caso la presenza di una struttura che presenta un RAMR oltre i limiti, ovvero l'ospedale di universitario di Brooklyn con un RAMR osservato pari a 7.

Al contempo i grafici riportati mettono in evidenza anche la presenza di strutture particolarmente virtuose: ad esempio, nel periodo 1997-1999, l'ospedale St. Francis ha riportato un RAMR pari a 1.29 e un valore di E intorno al 40, posizionandosi vicino al limite inferiore più estremo. Osserviamo che altre strutture hanno fatto riscontrare valori di RAMR inferiori; l'azione del parametro di precisione attraverso i limiti di tolleranza ci permette però di riconoscerle come strutture sostanzialmente nella media, cadendo infatti all'interno dei limiti al 95% e al 99.8%.

Possiamo quindi apprezzare il fatto che questa procedura di presentazione dei risultati permette di evitare classificazioni fuorvianti delle strutture monitorate. I grafici consultabili in [17] a pg 10 e in [18] a pg 10 riportano per ogni ospedale l'RAMR osservato e il limite di confidenza per quest'ultimo (basato sul modello di regressione logistica). Questo tipo di grafico (grafico 'caterpillar') è sicuramente di più complessa interpretazione rispetto a quelli riportati più avanti. Un altro errore sarebbe quello di riportare gli ospedali ordinati rispetto all'RAMR osservato: abbiamo infatti visto che una simile classificazione non mette in evidenza il ruolo della precisione alla quale la misura è stata riportata, creando sostanzialmente un ordinamento poco affidabile tra le strutture.

Tabella 1.1: OMR, EMR e RAMR relativi agli ospedali dello stato di New York nel periodo 1996-1998 (vedi [17]).

Ospedale	Casi	Morti	OMR	EMR	RAMR
Albany Medical Center	834	11	1.32	1.57	1.81
Arnot-Ogden	106	5	4.72	1.62	6.27
Bellevue	55	4	7.27	1.23	12.76
Beth Israel	446	5	1.12	2.28	1.06
Buffalo General	1082	20	1.85	1.92	2.07
Columbia Presbyterian	599	10	1.67	2.04	1.76
Ellis Hospital	447	6	1.34	1.59	1.82
Erie County	173	2	1.16	1.59	1.56
LIJ Medical Center	423	10	2.36	1.82	2.79
Lenox Hill	553	23	4.16	2.73	3.28
Maimonides	979	30	3.06	2.58	2.56
Millard Fillmore	642	6	0.93	1.83	1.10
Montefiore-Einstein	305	6	1.97	2.02	2.10
Montefiore-Moses	451	5	1.11	2.39	1.00
Mount Sinai	433	18	4.16	2.73	3.28
NY Hospital-Queens	303	6	1.98	1.59	2.69
NYU Hospitals Center	463	10	2.16	2.26	2.06
North Shore	784	19	2.42	2.27	2.30
Rochester General	920	25	2.72	2.72	2.15
St. Elizabeth	294	2	0.68	1.30	1.13
St. Francis	1646	26	1.58	2.10	1.62
St. Josephs	824	7	0.85	1.98	0.93
St. Lukes-Roosevelt	300	14	4.67	3.12	3.22
St. Peters	584	13	2.23	1.57	3.05
St. Vincents	551	25	4.54	1.81	5.39
Strong Memorial	347	9	2.59	2.41	2.32
United Health Services	331	8	2.42	2.32	2.25
Univ.Hosp.-Stony Brook	762	15	1.97	1.98	2.14
Univ.Hosp.-Upstate	523	15	2.87	2.35	2.63
Univ.Hosp. of Brooklyn	183	5	2.73	2.22	2.65
Weill Cornell-NYP	861	20	2.32	2.27	2.20
Westchester Medical Center	866	17	1.96	2.13	1.98
Winthrop Univ.Hosp.	744	8	1.08	2.83	0.82
Totale	18814	405	2.15		

Tabella 1.2: OMR, EMR e RAMR relativi agli ospedali dell stato di New York nel periodo 1997-1999 (vedi [18]).

Ospedale	Casi	Morti	OMR	EMR	RAMR
Albany Medical Center	759	16	2.11	1.69	2.79
Arnot-Ogden	88	3	3.41	1.52	5.03
Bellevue	80	2	2.50	1.86	3.01
Beth Israel	381	9	2.36	2.15	2.47
Buffalo General	1042	32	3.07	2.17	3.16
Columbia Presbyterian	553	13	2.35	1.91	2.76
Ellis Hospital	452	9	1.99	1.99	2.25
Erie County	106	3	2.83	1.56	4.06
LIJ Medical Center	375	3	0.80	1.83	0.98
Lenox Hill	612	14	2.29	2.15	2.38
Mainmonides	930	34	3.66	3.11	2.64
Millard Fillmore	748	19	2.54	1.89	3.01
Montefiore-Einstein	278	9	3.24	1.91	3.80
Montefiore-Moses	376	4	1.06	2.11	1.13
Mount Sinai	410	7	1.71	2.36	1.62
NY Hospital-Queen	280	3	1.07	1.87	1.28
NYU Hospitals Center	366	10	2.73	3.02	2.03
North Shore	816	11	1.35	1.84	1.65
Rochester General	874	16	1.83	2.79	1.47
St. Elizabeth	352	4	1.14	1.69	1.51
St. Francis	1804	23	1.27	2.22	1.29
St. Joseph	724	10	1.38	2.35	1.32
St Lukes-Roosevelt	245	7	2.86	2.75	2.33
St. Peters	605	3	0.50	1.64	0.68
St. Vincents	545	20	3.67	2.52	3.27
Strong Memorial	329	13	3.95	2.13	4.16
United Health Services	320	9	2.81	2.43	2.59
Univ.Hosp.-Stony Brook	719	13	1.81	2.01	2.02
Univ.Hosp.-Upstate	473	8	1.69	2.30	1.64
Univ.Hosp. of Brooklyn	284	22	7.75	2.48	7.00
Weill Cornell-NYP	704	14	1.99	2.64	1.69
Westchester Medical Center	816	29	3.55	2.57	3.10
Winthrop Univ.Hosp.	670	14	2.09	2.46	1.91
Totale	18116	406	2.24		

1.3.1 Un'alternativa per il parametro di precisione

Un'alternativa al numero di morti previste come parametro di precisione è la numerosità dei casi trattati dalla struttura. In [13] l'autore presenta infatti i medesimi grafici che riportiamo di seguito in funzione del volume di casi trattati invece che del volume di morti attese. Per fare ciò, conoscendo i limiti di controllo come funzione di E e non di N , è necessario riscalarne l'asse delle ascisse esprimendo E in funzione di N .

Considerando l'EMR medio pesato per le diverse numerosità, possiamo scrivere:

$$E = \frac{N \times \overline{EMR}}{100}.$$

È quindi possibile sostituire nell'espressione dei limiti di tolleranza il legame tra E ed N e plottare il grafico in funzione della numerosità trattata.

In Figura 1.5 riportiamo i Funnel Plot per i dati relativi al periodo 1997-1999, con limiti normali ed esatti. Le conclusioni e le riflessioni che possiamo trarre da quest'altro grafico sono sostanzialmente equivalenti a quelle che abbiamo ricavato dai grafici precedenti. Da un punto di vista intuitivo, una tale rappresentazione può apparire di più semplice ed immediata comprensione: la numerosità dei pazienti trattati dall'ospedale è chiaramente una misura di ρ meno criptica rispetto al numero atteso di morti. In ogni caso, per il modello che abbiamo utilizzato e per l'indice di performance scelto, la stima del numero atteso di morti (e quindi dell'EMR medio) è necessaria per disegnare correttamente i limiti di tolleranza.

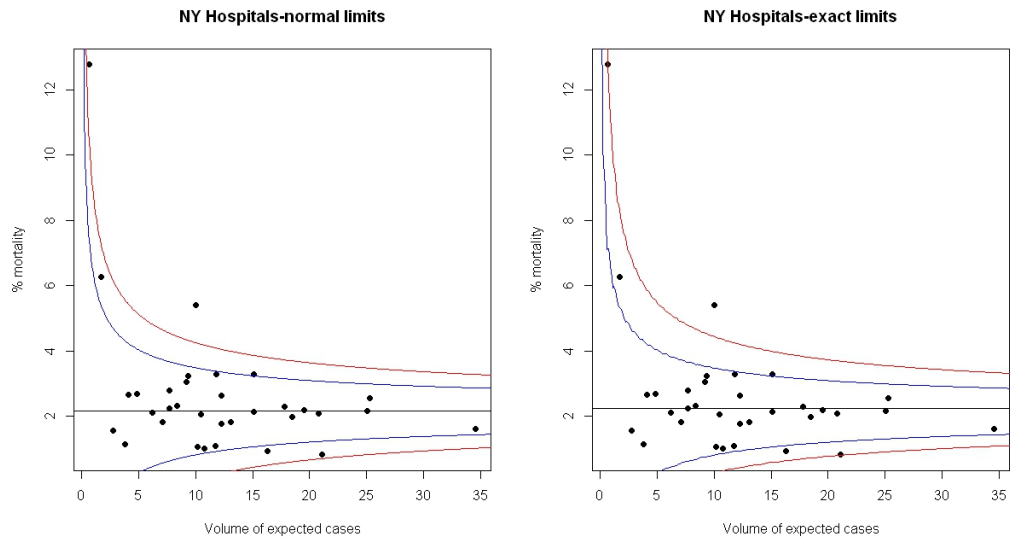


Figura 1.2: Funnel plot per i dati contenuti in Tabella 1.1. Nel primo grafico vi sono i limiti approssimati, nel secondo grafico i limiti esatti. In blu i limiti al 95%, in rosso i limiti al 99.8%.

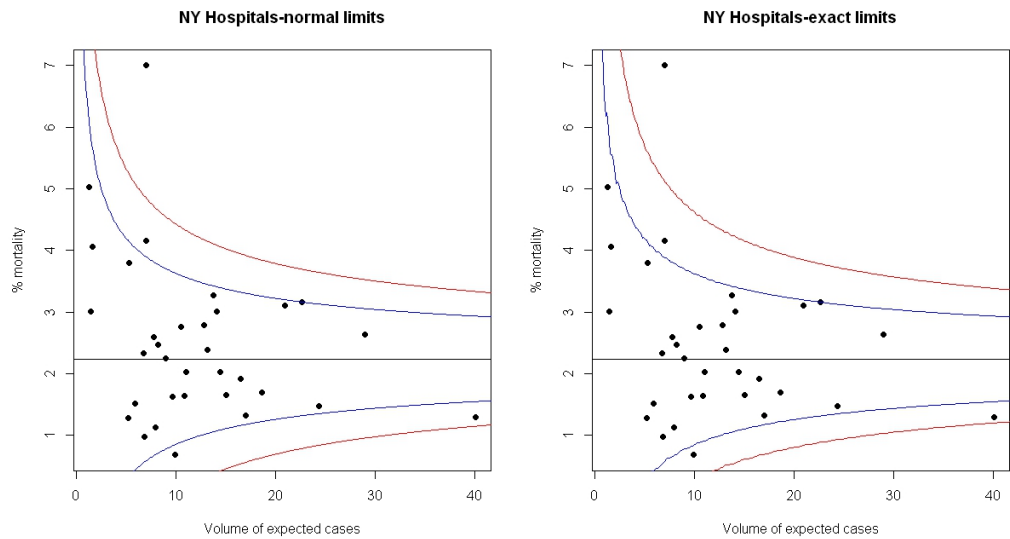


Figura 1.3: Funnel plot per i dati contenuti in Tabella 1.2. Nel primo grafico vi sono i limiti approssimati, nel secondo grafico i limiti esatti. In blu i limiti al 95%, in rosso i limiti al 99.8%.

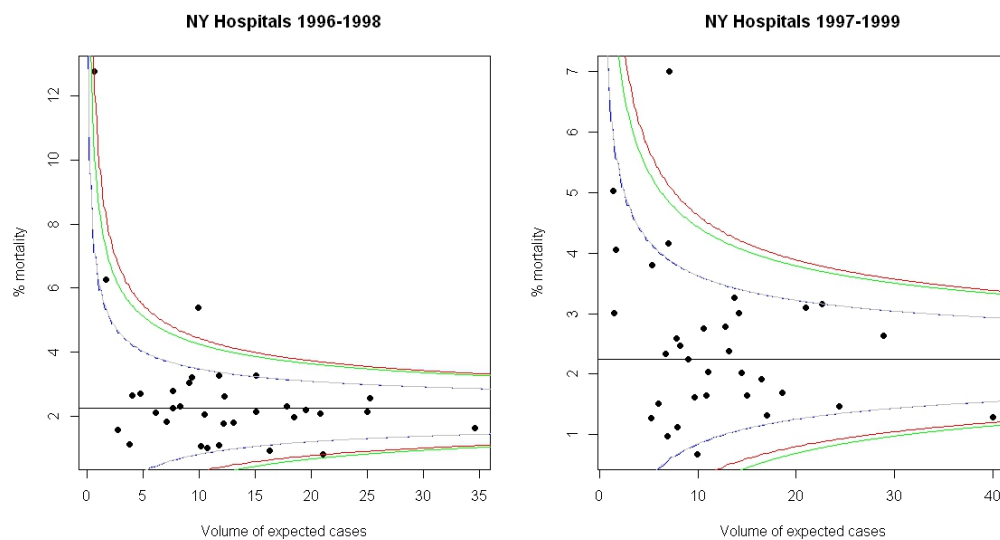


Figura 1.4: Limiti di tolleranza esatti (in blu e rosso) e approssimati (in grigio e verde) per i dati di Tabella 1.1 a sinistra e per i dati in Tabella 1.2 a destra.

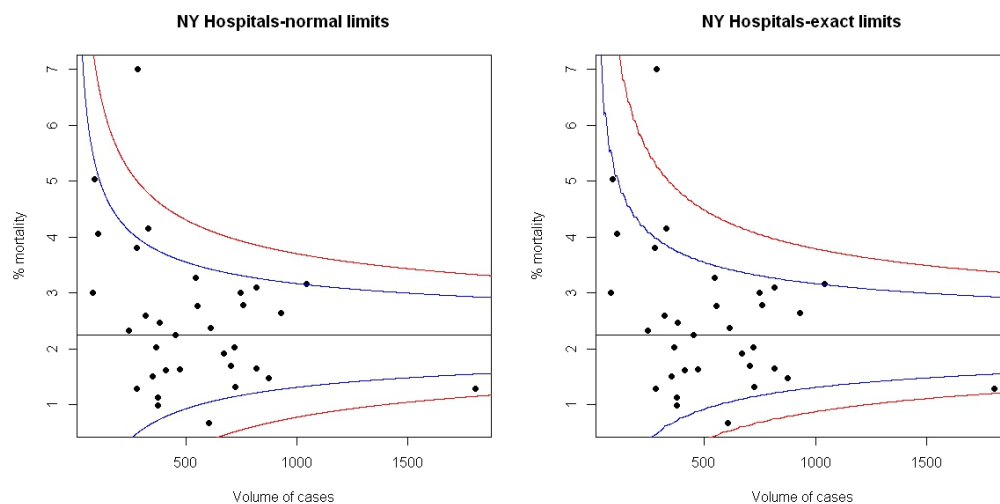


Figura 1.5: Funnel Plot per i dati contenuti in Tabella 1.2 costruito rispetto alla numerosità dei casi trattati.

Capitolo 2

Over-Dispersion

NEL precedente capitolo abbiamo introdotto i Funnel Plot discutendo la loro utilità e convenienza e convincendoci che in numerose situazioni si rivelano migliori di altri sistemi di comunicazione finale dei risultati. Abbiamo poi analizzato alcuni casi interessanti e ci siamo concentrati su un esempio, che ci ha permesso di mettere all'opera le potenzialità di questi strumenti.

Quello che può accadere però quando si ha a che fare con insiemi di dati più complessi di quelli utilizzati in precedenza è una esagerata dispersione delle performance intorno al target: molte unità studiate (spesso la maggior parte) possono risultare all'esterno dei limiti di controllo. I grafici che abbiamo considerato nel precedente capitolo erano infatti basati su un'ipotesi che è stata fin ora sottintesa: che la distribuzione che modella i dati in condizione di regolarità (di target) fosse in grado di spiegare adeguatamente la variabilità delle unità considerate sotto controllo. Questa ipotesi in molti casi è però poco verosimile; di conseguenza molte unità che dovrebbero essere giudicate come sotto controllo, in realtà appaiono maggiormente disperse (*over-dispersed*) intorno al target. Tutto ciò potrebbe falsare i nostri giudizi, inducendoci a considerare fuori controllo strutture le cui performance sono in realtà 'nella media', e portandoci in effetti ad una vera e propria contraddizione in termini: si studia cioè un fenomeno in cui la maggior parte degli eventi non è sotto controllo. Questo fenomeno di *over-dispersion* ci induce quindi a nutrire dubbi sulla validità dei limiti che abbiamo costruito. In altre parole, il modello scelto sotto l'ipotesi nulla non è adeguato e deve essere di conseguenza adattato.

In generale questo fenomeno è essenzialmente dovuto all'impatto di fattori che non sono stati fin ora correttamente considerati nella valutazione della variabilità: spesso non è l'impatto del singolo fattore a generare questo effetto indesiderato, ma il sommarsi di piccoli effetti dovuti a molti fattori

singolarmente poco incisivi sul fenomeno nel suo complesso.

È molto probabile che simili problemi insorgano maggiormente per quegli *outcome* che derivano da campioni molto numerosi (ad esempio ospedali che trattano decine o centinaia di migliaia di casi). Se la numerosità è utilizzata come indice della precisione come in molti dei casi illustrati precedentemente, al suo crescere i limiti di tolleranza tendono a diventare molto stretti; ma contemporaneamente aumentando le dimensioni dei campioni considerati, potrebbero entrare in gioco altri fattori di disturbo difficili da tenere sotto controllo.

Si rivela in questi casi necessario produrre degli aggiustamenti ai limiti di tolleranza al fine di eliminare questo fenomeno di sovra-dispersione. Le possibilità sono molteplici:

1. Valutare la possibilità di cambiare indicatore.
2. Misurare ulteriori fattori che contribuiscono alla variabilità eccessiva, oppure tentare una clusterizzazione delle strutture studiate in gruppi maggiormente omogenei (vedi [15]).
3. Stimare un fattore di over-dispersion, nel tentativo di espandere la variabilità del Funnel Plot prodotto.
4. Utilizzare un modello gerarchico (o *random effect model*) per caratterizzare l'indicatore di performance.

In questo capitolo vediamo come è possibile aggiustare il Funnel Plot seguendo il terzo punto appena presentato. Mostriamo in particolare due approcci semplici: uno moltiplicativo e uno additivo, basandoci sulle idee introdotte in [13] e in [15].

Nel prossimo capitolo vedremo invece come utilizzare un modello gerarchico e come riconoscere *outlier* ed *estremi*, di cui daremo una definizione, rispetto a quest'ultimo. Vedremo anche ripresentarsi sotto diversa luce l'approccio additivo che mostriamo più avanti.

Da questo punto in avanti considereremo che, sotto l'ipotesi di assenza di over-dispersion, valga $\rho = g(\theta_0)/\mathbb{V}_0(Y|\theta_0, \rho)$, dove con \mathbb{V}_0 indichiamo la varianza dell'indicatore non soggetta a sovra-dispersione (o *varianza nulla*).

2.1 Approccio moltiplicativo

Consideriamo un fattore moltiplicativo ϕ , che ha lo scopo di accrescere la varianza nulla nel seguente modo:

$$\mathbb{V}(Y|\theta_0, \rho, \phi) = \phi \mathbb{V}_0(Y|\theta_0, \rho) = \frac{\phi g(\theta_0)}{\rho} \quad (2.1)$$

Supponiamo di avere a disposizione I unità, e assumiamo che siano tutte considerabili sotto controllo. Da [7] e da [13] ci viene messa a disposizione una stima di ϕ :

$$\hat{\phi} = \frac{1}{I} \sum_{i=1}^I \frac{(y_i - \theta_0)^2 \rho_i}{g(\theta_0)} = \frac{1}{I} \sum_{i=1}^I z_i^2 \quad (2.2)$$

dove gli z_i sono i residui di Pearson (o *Z-scores*) definiti come:

$$z_i = \frac{y_i - \theta_0}{\sqrt{\mathbb{V}_0(Y|\theta_0)}}.$$

$\hat{\phi}$ altro non è quindi che una stima della varianza degli *Z-scores*.

Introducendo il parametro ϕ , i limiti di controllo possono essere riscritti aumentandone l'ampiezza di un fattore $\sqrt{\hat{\phi}}$. Per esempio nel caso di limiti normali abbiamo:

$$y_p(\theta_0, \rho) = \theta_0 + z_p \sqrt{\hat{\phi} \frac{g(\theta_0)}{\rho}}. \quad (2.3)$$

Vi è ancora un problema, ovvero quali unità in esame considerare come sotto controllo e di conseguenza quali unità utilizzare per la stima di ϕ . Includendo infatti nella stima anche le unità estreme, si andrebbe incontro ad un aumento del valore di $\hat{\phi}$ che determina un aumento ulteriore dei limiti di controllo, rendendo più difficoltoso riconoscere le strutture cui siamo realmente interessati, cioè quelle con performance inusuali. Il nostro scopo è quindi quello di rendere il più possibile robusta la stima di ϕ minimizzando l'influenza delle strutture estreme.

2.1.1 Stima *Winsorised*

In [13] viene proposta la seguente procedura di *Winsorization*:

1. I casi vengono ordinati in relazione al valore del relativo residuo Z .

2. Fissato un valore $q \in (0, 1)$, si identificano i valori Z_q e Z_{1-q} , ovvero i valori (esatti o approssimati) che lasciano alla propria sinistra il $100q\%$ e il $100(1 - q)\%$ degli Z_i .
3. I valori minori di Z_q vengono fissati pari a Z_q , mentre i valori maggiori di Z_{1-q} vengono fissati pari a Z_{1-q} . I nuovi Z-scores così modificati sono denotati con il simbolo $Z_i^W(q)$.
4. Possiamo calcolare la stima di $\hat{\phi}$ sostituendo gli $Z_i^W(q)$ nella formula (2.2):

$$\hat{\phi} = \frac{1}{I} \sum_{i=1}^I Z_i^W(q)^2.$$

Osserviamo che se non vi è presenza di over-dispersion, allora la quantità $I\hat{\phi}$ si distribuisce come una variabile aleatoria χ_I^2 , in quanto somma di quadrati di variabili aleatorie $N(0, 1)$. Ciò significa che $\mathbb{E}(\hat{\phi}) = 1$ e $\mathbb{V}(\hat{\phi}) = 2/I$. Ecco che, piuttosto che assumere direttamente l'over-dispersion, è meglio quindi agire come segue:

1. Non assumere l'*under-dispersion*; quindi se $\hat{\phi} < 1$, si può assumere $\phi = 1$.
2. Assumere l'over-dispersion solo in presenza di un valore 'statisticamente significativo' di $\hat{\phi}$; ad esempio assumeremo $\phi \neq 1$ se $\hat{\phi} > 1 + 2\sqrt{2/I}$.

Come esempio, riprendiamo i dati relativi al programma *CABG*, studiato negli ospedali dello stato di New York nel triennio 1997-1999.

In Figura 2.1 riportiamo il Funnel Plot che abbiamo già inserito precedentemente¹ e i Funnel Plot corretti secondo il metodo moltiplicativo con $q = 0.05$ e $q = 0.1$. Nel primo caso si ottiene un valore di $\hat{\phi} = 1.567$ mentre nel secondo caso si ha $\hat{\phi} = 1.396$. Nel primo caso si ha

$$\hat{\phi} > 1 + 2\sqrt{\frac{2}{I}} = 1 + 2\sqrt{\frac{2}{33}} = 1.49$$

quindi saremmo guidati ad inserire il termine moltiplicativo. Nel secondo caso sembra invece che si possa trascurare tranquillamente la sovra-dispersione.

Osservando i grafici ottenuti (abbiamo comunque riportato anche il grafico per $q = 0.1$), le conclusioni cui siamo portati sono sostanzialmente equivalenti a quelle espresse nel Capitolo 1. Per questo esempio quindi l'eventuale sovra-dispersione non è particolarmente incisiva, neanche per $q = 0.05$.

¹Vedi Figura 1.3.

2.1.2 Selezione diretta

Un'altra possibilità può essere quella di selezionare direttamente le strutture da utilizzare per stimare il parametro di aggiustamento. Proviamo ad osservare direttamente i valori degli Z-scores. In Figura 2.2 a sinistra è riportato il Q-Q Plot relativo a tutti Z-scores; si osserva chiaramente la presenza di un valore non allineato sulla coda a destra del grafico. Osservando la situazione, possiamo provare ad escludere questo ospedale (l'ospedale Universitario di Brooklyn) dalla stima di ϕ . Osserviamo in Figura 2.2 a destra il Q-Q Plot degli Z-scores privati dello score più estremo: la situazione sembra ora più omogenea.

Utilizzando le rimanenti trentadue strutture per stimare il parametro, abbiamo $\hat{\phi} = 1.657$. Risulta in questo caso:

$$\hat{\phi} > 1 + 2\sqrt{\frac{2}{32}} = 1.5$$

Anche se non di molto, il valore stimato supera la soglia consigliata.

In Figura 2.3 sono affiancati i Funnel Plot senza e con la correzione appena calcolata. Anche in questo le conclusioni cui siamo indotti non si modificano introducendo l'over-dispersion nel nostro grafico.

Facciamo notare però che potrebbe rivelarsi complesso scegliere 'a occhio' le strutture da utilizzare nella stima². Il procedimento di *Winsorization* mostrato precedentemente è quindi più consigliabile.

2.2 Approccio additivo

Consideriamo un approccio alternativo, utilizzando un modello del tipo *random effect*: assumiamo che $Y_i|\theta_i$ abbia valore atteso pari a θ_i e varianza pari a σ_i^2 , e che per le unità sotto controllo, θ_i sia distribuito con valore atteso θ_0 e varianza τ^2 . Vedremo successivamente più in dettaglio nel Capitolo 3 che una stima di τ^2 viene data da:

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I - 1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i} \quad (2.4)$$

dove $w_i = 1/\sigma_i^2$; $\hat{\phi}$ è calcolata come in precedenza ed è possibile fare la seguente verifica di omogeneità: se $\hat{\phi} < (I - 1)/I$ allora è consigliabile assumere

²Addirittura la quasi totalità delle unità studiate potrebbe essere da scartare con un procedimento simile.

$\tau = 0$, ovvero la completa omogeneità. I limiti corretti diventano quindi:

$$y_p = \theta_0 + z_p \sqrt{\mathbb{V}_0(Y|\theta_0, \rho) + \hat{\tau}^2} = \theta_0 + z_p \sqrt{\frac{g(\theta_0)}{\rho} + \hat{\tau}^2}. \quad (2.5)$$

Quello che accade è che, al crescere della numerosità (o del numero atteso di eventi se quest'ultimo è utilizzato come indice di precisione), i limiti di tolleranza avranno un'ampiezza che non tenderà più a zero come in precedenza, ma ad un valore limite, dato da $z_p \tau$. Si tiene quindi conto di eventuali effetti che insorgono ad alte numerosità.

Come esempio utilizziamo i dati presentati nel Capitolo 1 in Tabella 1.1, relativi al progetto *CABG* condotto negli ospedali dello stato di New York nel triennio 1996-1998. Per il calcolo di $\hat{\phi}$, utilizziamo la stima *Winsorised* con $q = 0.1$. I valori w_i sono stimati come $\frac{\rho}{g(\theta_0)}$. Si ottiene un valore di $\hat{\phi}$ pari a 1.33 (osserviamo che non sarebbe stato necessario applicare la correzione moltiplicativa dato il valore inferiore alla soglia consigliata) e di conseguenza avremo $\hat{\tau}^2 = 0.03$.

Quello che possiamo notare osservando i grafici riportati in Figura 2.4 è che l'imbuto formato dai limiti di controllo tende ad essere maggiormente aperto all'aumentare del parametro di precisione, mentre a basse numerosità di casi previsti i due plot sono praticamente identici. L'effetto sulla parte destra del grafico è comunque contenuto; anche per questi dati le nostre opinioni non vengono modificate dopo l'inserimento della correzione additiva.

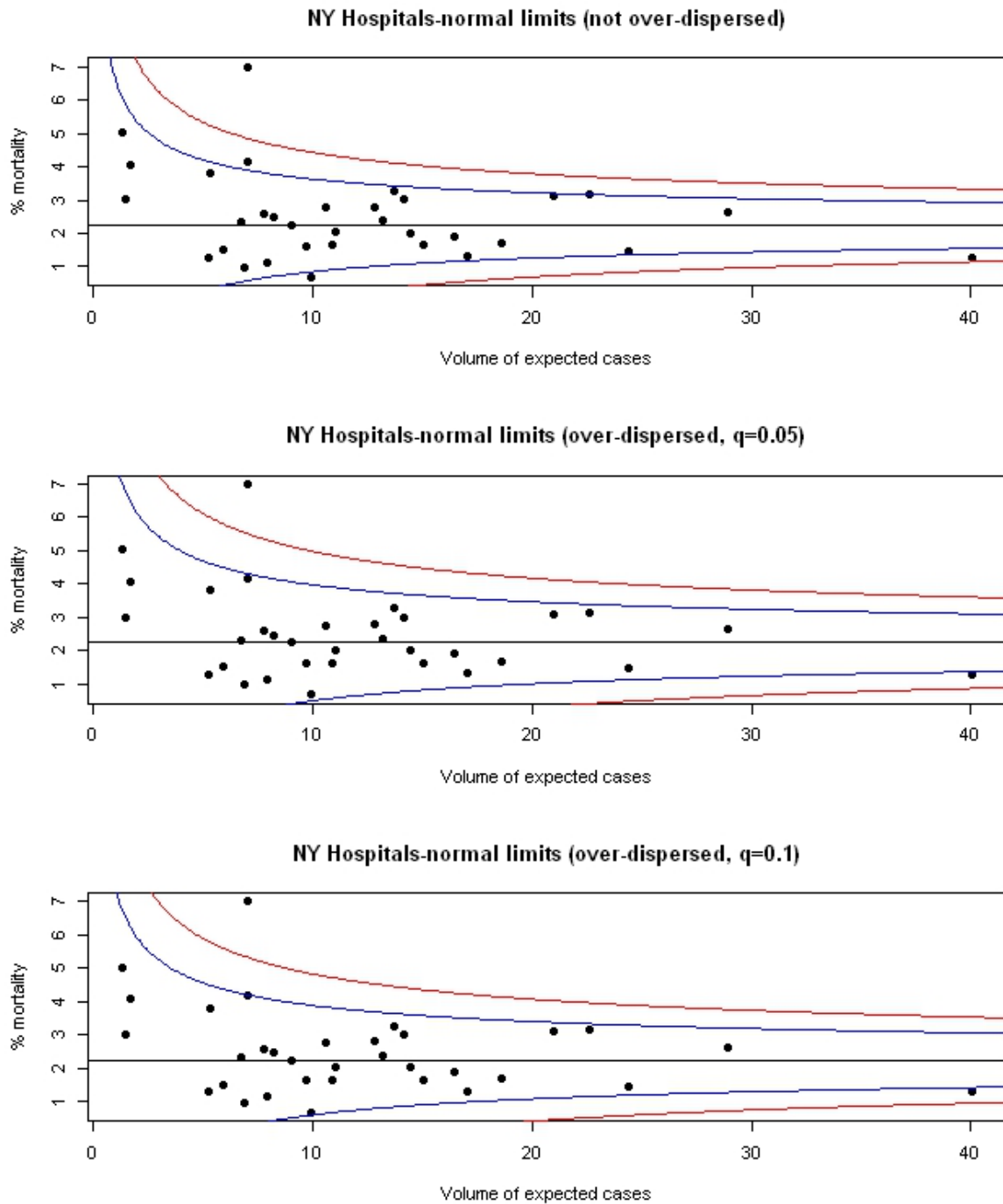


Figura 2.1: Dall'alto: Funnel Plot senza contributo moltiplicativo e Funnel Plots con contributo moltiplicativo ($q = 0,05$ e $q = 0,1$) per i dati contenuti in Tabella 1.2.

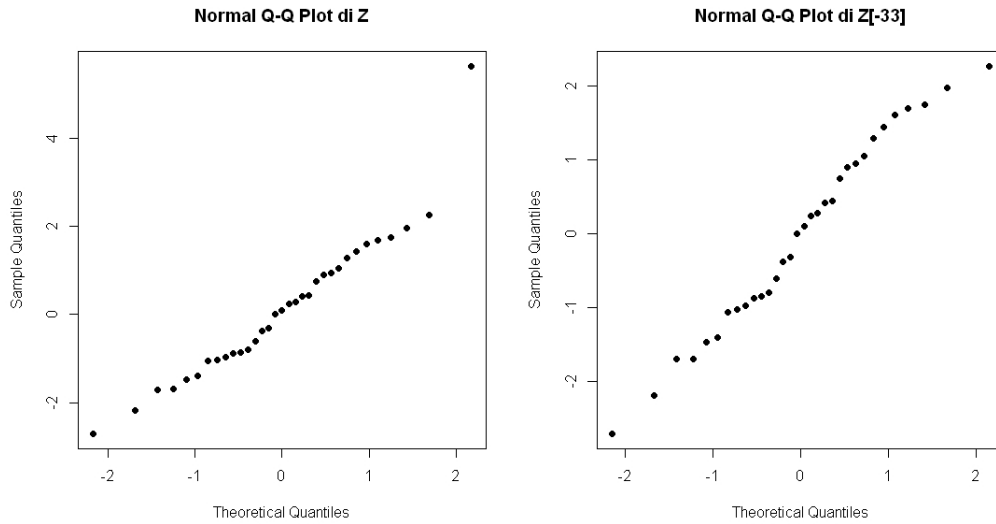


Figura 2.2: Q-Q Plot degli Z-scores relativi ai dati in Tabella 1.2. Nel secondo grafico è stato eliminato l'ospedale di Brooklyn che, come si può osservare, è molto probabilmente una struttura fuori controllo.

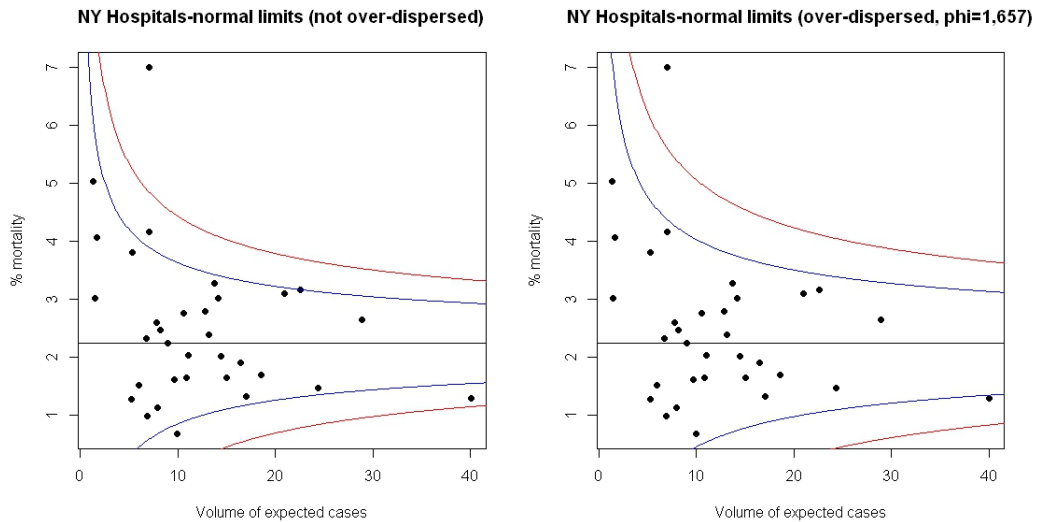


Figura 2.3: Da sinistra: Funnel Plot non corretto e Funnel Plot corretto (con $\phi = 1.657$) per i dati contenuti in Tabella 1.2.

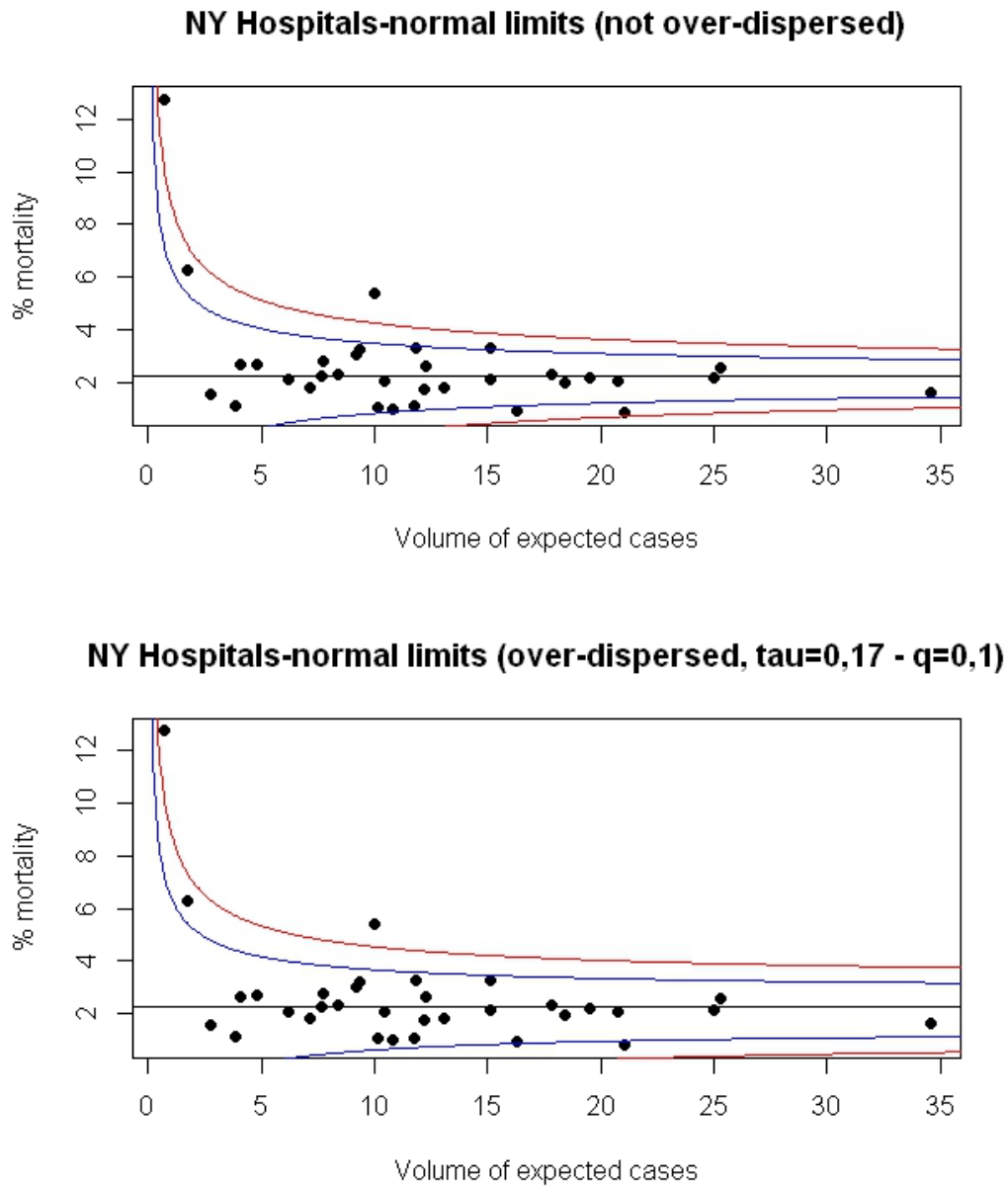


Figura 2.4: Dall'alto: Funnel Plot non corretto e Funnel Plot corretto con la costante additiva τ^2 per i dati di Tabella 1.1.

Capitolo 3

Utilizzo di Modelli Gerarchici

TUTT'OGGI è oramai consuetudine utilizzare modelli di tipo gerarchico, come accennato nel Capitolo 2, per fare confronti tra performance di strutture sanitarie (e non solo). Modelli di questo tipo, oltre ad essere maneggevoli e facilmente adattabili a molte situazioni, permettono di assumere l'esistenza di fattori casuali specifici per ogni *health-provider*, i quali effetti vengono considerati provenienti da una distribuzione comune.

L'utilizzo di tali modelli diventa quindi uno strumento molto potente ad esempio quando si ha a che fare con il problema dell'over-dispersion (introdotto nel Capitolo 2) nella costruzione di Funnel Plot; utilizzando modelli più raffinati, è possibile infatti tenere conto di eventuali fattori sfuggiti in modellazioni più semplicistiche (come quelle presentate ad esempio nel Capitolo 1).

Vogliamo quindi presentare in questo capitolo alcuni approcci che sfruttano la costruzione di modelli gerarchici al fine di riconoscere le strutture che presentano performance inusuali: come viene rimarcato in [6], faremo una distinzione tra *outlier* e *estremi*:

- Per *outlier* intendiamo una performance osservata che giace oltre il limite posto dal modello assunto, che non è spiegata dal modello e che quindi non può essere considerata sotto controllo.
- Per *estremo* intendiamo una performance osservata che può essere spiegata dal modello assunto, ma che assume un valore inusuale rispetto alla distribuzione comune dell'effetto casuale.

In questo ambiente i Funnel Plot presentati nei Capitoli 1 e 2 diventano importanti strumenti di comunicazione grafica dei risultati dell'analisi condotta. Vedremo nel seguito quali informazioni possono darci tali Funnel Plot, in cui i limiti di confidenza verranno costruiti utilizzando diverse misure della variabilità tratte dal modello.

Per semplicità, in questo capitolo adottiamo un modello gerarchico di tipo Gaussiano a due livelli (*random effect model*), ma modelli differenti e più complessi potrebbero essere utilizzati in alternativa. Ne presentiamo gli aspetti principali qui di seguito.

3.1 *Random effect model*

Come già in precedenza, denotiamo con Y_i una misura di performance per l' i -esima struttura sotto indagine, e con I il numero totale di strutture studiate. Il modello che consideriamo è il seguente:

$$Y_i | \theta_i \stackrel{ind}{\sim} N(\theta_i, \sigma_i^2), \quad \theta_i \stackrel{iid}{\sim} N(\mu, \tau^2) \quad (3.1)$$

dove le σ_i^2 sono assunte come note¹. Stiamo cioè assumendo che ogni unità abbia un proprio valore atteso, il quale a sua volta è proveniente da una distribuzione comune intorno ad una media generale μ e con deviazione standard τ . Si introduce quindi una ulteriore parte di variabilità dovuta alle variazioni *tra* strutture. Le Y_i sono indipendenti ma chiaramente non identicamente distribuite, come invece sono i θ_i .

Vediamo le principali caratteristiche di questo modello. Dal punto di vista Bayesiano si ha quanto segue:

Proposizione 3.1. *Dato il modello precedentemente descritto vale:*

$$\theta_i | y_i, \mu, \tau \sim N(\hat{\theta}_i, a_i \sigma_i^2)$$

dove

$$\hat{\theta}_i = a_i y_i + (1 - a_i) \mu, \quad a_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}. \quad (3.2)$$

Dimostrazione. Per ricavare la densità a posteriori dei θ_i , utilizziamo il teorema di Bayes e l'indipendenza:

¹Per le σ_i^2 può essere nota ad esempio una relazione in funzione di un parametro di precisione ρ .

$$\begin{aligned}
g(\theta_i|y_i) &\propto f(y_i|\theta_i) \times g(\theta_i) \\
&\propto \exp\left\{\frac{1}{2}\left\{\frac{(y_i - \theta_i)^2}{\sigma_i^2} + \frac{(\theta_i - \mu)^2}{\tau^2}\right\}\right\} \\
&\propto \exp\left\{\frac{1}{2}\left\{\frac{y_i^2 - 2y_i\theta_i + \theta_i^2}{\sigma_i^2} + \frac{\theta_i^2 - 2\mu\theta_i + \mu^2}{\tau^2}\right\}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left\{\frac{\tau^2 y_i^2 - 2\tau^2 y_i \theta_i + \tau^2 \theta_i^2 + \sigma_i^2 \theta_i^2 - 2\mu\theta_i \sigma_i^2 + \sigma_i^2 \mu^2}{\sigma_i^2 \tau^2}\right\}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_i^2 \tau^2}\left\{\theta_i^2(\tau^2 + \sigma_i^2) - 2\theta_i(\tau^2 y_i + \sigma_i^2 \mu + \tau^2 y_i^2 + \sigma_i^2 \mu^2)\right\}\right\} \\
&\propto \exp\left\{-\frac{\tau^2 + \sigma_i^2}{2\sigma_i^2 \tau^2}\left\{\theta_i - \frac{\tau^2 y_i + \sigma_i^2 \mu}{\tau^2 + \sigma_i^2}\right\}^2\right\} \times \\
&\quad \times \exp\left\{-\frac{\tau^2 + \sigma_i^2}{2\sigma_i^2 \tau^2}\left\{\frac{\tau^2 y_i^2 + \sigma_i^2 \mu^2}{\tau^2 + \sigma_i^2} - \frac{(\tau^2 y_i + \sigma_i^2 \mu)}{(\tau^2 + \sigma_i^2)^2}\right\}\right\}
\end{aligned}$$

Il secondo esponenziale nell'ultima espressione non dipende da θ_i , di conseguenza sarà inglobato dalla costante di normalizzazione.

Consideriamo solamente quindi la seguente espressione:

$$\exp\left\{-\frac{\tau^2 + \sigma_i^2}{2\sigma_i^2 \tau^2}\left\{\theta_i - \frac{\tau^2 y_i + \sigma_i^2 \mu}{\tau^2 + \sigma_i^2}\right\}^2\right\}$$

nella quale riconosciamo il nucleo di una densità Gaussiana con la media e la varianza riportate in (3.2). □

La stima $\hat{\theta}_i$ altro non è che la stima di *shrinkage* del parametro θ_i . Come spesso accade nell'analisi Bayesiana, a seconda del valore di un peso, in questo caso a_i , si dà diversa preponderanza alle due componenti della stima, quella a priori nel caso a_i sia piccolo, o ai dati a disposizione se a_i è grande. Il valore di a_i per il singolo ospedale sotto indagine dipende da molte caratteristiche, ma in principal modo dal valore di σ_i^2 , che è legato alla precisione con cui l'informazione sulla performance è stata ricevuta. Un ospedale che tratta pochi casi ha probabilmente una bassa precisione e quindi il suo indicatore Y_i presenterà una maggiore incertezza, dando maggior peso all'informazione a priori; viceversa, nel caso di ospedali molto più grandi che quindi trattano quantità di casi più estese e che presenteranno degli indicatori Y_i affetti da deviazione standard inferiore, si darà più peso all'informazione proveniente dai dati. Anche τ^2 ha un effetto da non trascurare: se τ^2 ha un valore elevato

rispetto a σ_i^2 , può mascherare l'effetto di quest'ultima, rendendo a_i vicino a uno.

Adottando un approccio più consueto, chiamato *empirical Bayes*, i valori di μ e τ^2 vengono stimati dalla distribuzione marginale dei dati. Vale infatti quanto segue:

Proposizione 3.2. *Dato il modello precedentemente descritto vale:*

$$Y_i \stackrel{ind}{\sim} N(\mu, \sigma_i^2 + \tau^2).$$

Dimostrazione. Per calcolare la marginale dei dati Y_i è necessario calcolare il seguente integrale:

$$f(\mathbf{y}) = \int_{\mathbb{R}^I} \left(\prod_{i=1}^I \frac{1}{\sqrt{2\pi\sigma_i^2}} \frac{1}{\sqrt{2\pi\tau^2}} \exp \left\{ -\frac{1}{2} \left\{ \frac{(y_i - \theta_i)^2}{\sigma_i^2} + \frac{(\theta_i - \mu)^2}{\tau^2} \right\} \right\} \right) d\theta_1 \cdots d\theta_I.$$

Si può osservare però che l'integranda è fattorizzabile; di conseguenza con un leggero abuso di linguaggio possiamo scrivere:

$$f(\mathbf{y}) = \prod_{i=1}^I f(y_i). \quad (3.3)$$

Che corrisponde a dire che, marginalmente, le osservazioni Y_1 saranno indipendenti. Risolviamo quindi il seguente integrale per la sola y_1 :

$$\begin{aligned} f(y_1) &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{\sigma_1^2\tau^2}} \exp \left\{ -\frac{1}{2} \left\{ \frac{y_1^2 - 2y_1\theta_1 + \theta_1^2}{\sigma_1^2} + \frac{\theta_1^2 - 2\mu\theta_1 + \mu^2}{\tau^2} \right\} \right\} d\theta_1 \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\tau^2}} \int_{-\infty}^{+\infty} \exp \left\{ -\frac{1}{2} \left\{ \frac{\tau^2 y_1^2 - 2\tau^2 y_1 \theta_1 + \tau^2 \theta_1^2 + \sigma_1^2 \theta_1^2 - 2\mu\theta_1 \sigma_1^2 + \sigma_1^2 \mu^2}{\sigma_1^2 \tau^2} \right\} \right\} d\theta_1 \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\tau^2}} \int_{-\infty}^{+\infty} \exp \left\{ -\frac{1}{2} \left\{ \frac{\theta_1^2 (\tau^2 + \sigma_1^2) - 2\theta_1 (\tau^2 y_1 + \sigma_1^2 \mu + \tau^2 y_1^2 + \sigma_1^2 \mu^2)}{\sigma_1^2 \tau^2} \right\} \right\} d\theta_1 \end{aligned}$$

Riarrangiando i termini nell'esponenziale si arriva a scrivere

$$\frac{\exp \left\{ -\frac{\tau^2 + \sigma_1^2}{2\sigma_1^2 \tau^2} \left\{ \frac{\tau^2 y_1^2 + \sigma_1^2 \mu^2}{\tau^2 + \sigma_1^2} - \frac{(\tau^2 y_1 + \sigma_1^2 \mu)^2}{(\tau^2 + \sigma_1^2)^2} \right\} \right\}}{2\pi\sqrt{\sigma_1^2\tau^2}} \int_{-\infty}^{+\infty} \exp \left\{ -\frac{\tau^2 + \sigma_1^2}{2\sigma_1^2 \tau^2} \left\{ \theta_1 - \frac{\tau^2 y_1 + \sigma_1^2 \mu}{\tau^2 + \sigma_1^2} \right\}^2 \right\} d\theta_1.$$

Nell'integrale ora si può riconoscere il nucleo di una densità Gaussiana rispetto a θ_1 ; di conseguenza abbiamo:

$$f(y_1) = \frac{1}{2\pi\sqrt{\sigma_1^2\tau^2}} \sqrt{2\pi \frac{\sigma_1^2\tau^2}{\tau^2 + \sigma_1^2}} \exp\left\{-\frac{\tau^2 + \sigma_1^2}{2\sigma_1^2\tau^2} \left\{ \frac{\tau^2 y_1^2 + \sigma_1^2 \mu^2}{\tau^2 + \sigma_2} - \frac{(\tau^2 y_1 + \sigma_1^2 \mu)^2}{(\tau^2 + \sigma_2)^2} \right\}\right\}$$

Svolgendo i conti e semplificando, otteniamo:

$$\frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2)}} \exp\left\{-\frac{1}{2} \frac{(y_1 - \mu)^2}{\tau^2 + \sigma_1^2}\right\}$$

che insieme all'indipendenza già verificata sopra ci dà la tesi. \square

Indicando con w_i la quantità $1/\sigma_i^2$, abbiamo

$$\hat{\tau}^2 = \max\left(0, \frac{\sum_i w_i (y_i - y_w)^2 - (I - 1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}\right) \quad (3.4)$$

dove

$$y_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

mentre la stima di μ è data da

$$\hat{\mu} = \frac{\sum_i a_i y_i}{\sum_i a_i}$$

cioè una media pesata delle osservazioni.

Prima di proseguire, facciamo osseverare che il modello presentato è già stato utilizzato nel Capitolo 2 riguardo alla correzione additiva per l'overdispersion, con l'ipotesi aggiuntiva di Gaussianità. Infatti, nella marginale dei dati Y , compare il termine additivo τ^2 che rappresenta la varianza aggiuntiva dovuta ai θ_i .

La formula di stima utilizzata in quel caso era la seguente:

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I - 1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}.$$

Quest'ultima è equivalente alla (3.2). In particolare avevamo:

$$\begin{aligned}
 I\hat{\phi} &= \sum_{i=1}^I z_i^2 & (3.5) \\
 &= \sum_{i=1}^I \frac{(y_i - \theta_0)^2}{\sigma_i^2} \\
 &= \sum_{i=1}^I w_i (y_i - \theta_0)^2
 \end{aligned}$$

che equivale a $\sum_i w_i (y_i - y_w)^2$ nel caso in cui il valore di target per $\mathbb{E}(Y_i|\theta_0)$ sia imposto nella distribuzione 'nulla' (pari appunto a θ_0 , in questo caso equivalente a y_w).

Come già osservato nel Capitolo 2, nella pratica si desidera avere una stima robusta di τ^2 , in modo che quest'ultima non risenta in maniera troppo forte delle unità che sono effettivamente sovra-disperse. Per esempio si può utilizzare una stima *Winsorised* di Φ , come già fatto in precedenza.

I valori di w_i erano basati poi sull'ipotesi fatta sulla varianza dell'indice di performance, ovvero:

$$w_i = \frac{\rho}{g(\theta_0)}.$$

3.2 Tre possibili procedimenti

Presentiamo tre procedimenti per individuare i provider che presentano delle performance inusuali o notevoli; seguendo lo schema proposto in [6], questi sono nominati come segue:

- **Approccio 1:** basato sull'identificazione di outliers rispetto al modello *common mean*.
- **Approccio 2:** basato sull'identificazione di outliers rispetto al modello *random effect*.
- **Approccio 3:** basato sull'identificazione degli estremi rispetto al modello *random effect*.

Il primo approccio consiste nell'utilizzare un modello *common mean*, ovvero un modello non gerarchico, come ipotesi nulla e nel compiere dei test per verificare la validità di quest'ultimo. Esso è in effetti equivalente ai modelli già presentati nel Capitolo 1. Un modello di questo tipo è certamente molto

semplice, ma come abbiamo già fatto presente, vi sono molte ragioni a favore dell'utilizzo di un modello più accurato (quello appena presentato o altri più complessi). Questo porta a concentrare la nostra attenzione sugli approcci 2 e 3.

Il secondo metodo pone il modello ad effetto casuale sotto l'ipotesi nulla ed identifica eventuali deviazioni da quest'ultimo, mentre il terzo procedimento è basato sulle probabilità a posteriori degli effetti casuali.

La scelta dovrebbe essere basata sulla strategia di modellazione selezionata dall'analista:

1. *Hypothesis Testing Strategy*: Il modello random effect è posto nell'ipotesi nulla; vengono permesse alcune variazioni intorno alla media della popolazione, variazioni assunte Normali, e dovute a fattori incogniti che vanno oltre il controllo dei diversi provider. Questo approccio è consigliato quando si sta facendo ricerca di outlier del modello. È consigliabile procedere quindi con l'approccio 2.
2. *Estimation Strategy*: Il modello gerarchico viene costruito pensando di adattarlo a tutti i dati. Non è quindi sensato procedere alla ricerca degli outliers, dato che per costruzione non ve ne sono. È consigliabile quindi utilizzare il terzo approccio, legato alla ricerca di estremi.

È il caso di sottolineare come si intende riconoscere una performance estrema quando si utilizza il terzo metodo. Alcuni autori (vedi [11], [12], [4] e [16]) utilizzano metodi basati su una probabilità a posteriori della performance di essere maggiore della media della popolazione che sia sufficientemente grande. Come abbiamo già osservato indirettamente nei capitoli precedenti, non è necessario che una struttura inusuale sia 'estrema', ma è cruciale l'intervento della numerosità o di un'altro parametro di precisione per determinare la giusta scala su cui valutare ogni singola performance. Altri autori sottolineano questo fatto, e suggeriscono l'utilizzo di un target esterno per la valutazione delle strutture estreme (vedi [2], [3] e [9]).

Al di là di questo, il problema principale è che, quando vengono utilizzati modelli gerarchici, in letteratura spesso non è chiaro quale degli approcci presentati è usato, principalmente a causa del fatto che gli autori non sono chiari nello specificare quale standard error viene utilizzato per costruire i limiti di confidenza. Quello che vogliamo mostrare (e che viene sottolineato in [6]) è che la diversa scelta tra gli approcci proposti è legata alla scelta dell'errore standard da utilizzare.

Esploriamo di seguito nel dettaglio i tre procedimenti presentati brevemente qui sopra.

3.2.1 Approccio 1

Assumendo il modello common mean, stiamo sostanzialmente assumendo un modello non gerarchico. Ciò equivale ad assumere $\tau = 0$. Di conseguenza il nostro modello diventa $Y_i \sim N(\theta_i, \sigma_i^2)$, e il nostro interesse si focalizza su $H_0 : \theta_i = \mu$ contro l'ipotesi alternativa $H_1 : \theta_i > \mu$ (o viceversa $H_1 : \theta_i < \mu$), per ogni unità sotto esame.

La statistica test di interesse in un caso del genere è

$$Z_i^{(1)} = \frac{y_i - \mu}{\sigma_i}$$

ed il p-value relativo al test per ogni unità è:

$$p_i^{(1)} = 1 - \Phi(Z_i^{(1)}).$$

Osserviamo che il modello che stiamo assumendo corrisponde (nel particolare caso di distribuzione Gaussiana, quindi di limiti normali) a quelli che abbiamo visto e utilizzato nel Capitolo 1, e la statistica $Z_i^{(1)}$ corrisponde per ogni unità allo Z-score. Abbiamo poi utilizzato μ di fatto come valore di target; svolge quindi il ruolo svolto in precedenza da θ_0 .

Si potrebbe affrontare lo stesso problema da un punto di vista Bayesiano: il relativo p-value bayesiano che si otterrebbe è identico a $p_i^{(1)}$ (vedi [6] per i dettagli).

3.2.2 Approccio 2

Consideriamo l'espressione dei $\hat{\theta}_i$ data dalla formula (3.2): se θ_i proviene dalla stessa distribuzione degli altri provider, allora $\hat{\theta}_i$ è uno stimatore non distorto della media della popolazione μ . Altrimenti il valore atteso di $\hat{\theta}_i$ assumerà un valore differente, ad esempio μ_i .

Considerando quindi $\hat{\theta}_i$ come stima di μ_i , testiamo l'ipotesi nulla $H_0 : \mu_i = \mu$ contro l'ipotesi $H_1 : \mu_i > \mu$ (o viceversa $H_1 : \mu_i < \mu$) per ogni provider. Il rifiuto di H_0 porta a considerare θ_i troppo grande (o troppo piccola) per provenire dalla relativa distribuzione nulla. Quello che andiamo quindi a fare è cercare quelle unità che verosimilmente provengono da una distribuzione diversa dalla distribuzione nulla dei θ_i , ovvero che hanno una media μ_i che, rispetto alla variabilità consentita dal modello, sia troppo diversa dalla media comune.

Sotto H_0 , si può ricavare molto semplicemente che:

$$\hat{\theta}_i \sim N(\mu, a_i^2(\sigma_i^2 + \tau^2)).$$

In particolare:

$$\begin{aligned}\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i) &= a_i^2(\sigma_i^2 + \tau^2) \\ &= \frac{\tau^4}{\sigma_i^2 + \tau^2} \\ &= a_i\tau^2\end{aligned}$$

$\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i)$ rappresenta la varianza di θ_i dalla distribuzione congiunta di Y_i e θ_i sotto H_0 .

La statistica test diventa:

$$\begin{aligned}Z_i^{(2)} &= \frac{\hat{\theta}_i - \mu}{\sqrt{\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i)}} \\ &= \frac{a_i(y_i - \mu)}{\sqrt{a_i^2(\sigma_i^2 + \tau^2)}} \\ &= \frac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}.\end{aligned}$$

Il relativo p-value diventa:

$$p_i^{(2)} = 1 - \Phi(Z_i^{(2)}).$$

Come prima, applicando un approccio Bayesiano, il relativo p-value che si otterrebbe è identico a $p_i^{(2)}$ appena scritto (per i dettagli vedi [6]).

Facciamo una osservazione; questo secondo approccio è molto simile al precedente, ma vi è una differenza molto importante: viene permessa una ulteriore variabilità, dovuta al fatto che i θ_i possono avere delle oscillazioni intorno alla media comune μ (o θ_0 nei precedenti capitoli). Queste variazioni vengono permesse in ragione della costante τ^2 che va ad aggiungersi nel termine della varianza, rappresentando di conseguenza il contributo della variabilità intrinseca dei θ_i . Viene inclusa cioè nel modello una possibile over-dispersion di tipo additivo. I limiti prodotti con un simile approccio saranno in generale più larghi di quelli costruiti col precedente metodo, in particolare all'aumentare del parametro di precisione: se infatti la quantità σ_i^2 si riduce all'aumentare di ρ , il termine τ^2 conserva inalterata la sua influenza.

3.2.3 Approccio 3

Vogliamo ora concentrarci sui θ_i più che sulle μ_i , ovvero sulle realizzazioni dei *rate* di ogni unità. Consideriamo ancora i valori $\hat{\theta}_i$: se utilizzati come

stima di θ_i piuttosto che di μ_i , è necessario tenere conto della distorsione cui si va incontro utilizzando un'appropriata misura dell'errore.

In [5] e in [6] si distingue tra due misure alternative per lo standard error del fattore casuale:

1. standard error *diagnostico*,

$$\sqrt{\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i - \mu)} = \sqrt{\mathbb{V}(\hat{\theta}_i)},$$

che equivale per nostro modello a $\tau\sqrt{a_i}$.

2. standard error *comparativo*,

$$\sqrt{\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i - \theta_i)}.$$

Per il nostro modello:

$$\begin{aligned} \mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i - \theta_i) &= \mathbb{E}_{Y_i, \theta_i}(\hat{\theta}_i - \theta_i)^2 \\ &= \mathbb{E}_{Y_i}[\mathbb{E}_{\theta_i|Y_i}(\hat{\theta}_i - \theta_i)^2] \\ &= \mathbb{E}_{Y_i}[\mathbb{V}(\theta_i|Y_i)] \\ &= \sigma_i\sqrt{a_i}. \end{aligned}$$

Nell'ultimo passaggio, ricordiamo che $\mathbb{V}(\theta_i|Y_i)$ non dipende dai dati Y_i .

In [5] e in [6] si consiglia di utilizzare l'errore standard diagnostico per analizzare, ad esempio, le proprietà dei residui, nonché per la ricerca di outlier del modello. Possiamo quindi notare come non sia un caso che, nella precedente sezione riguardante il secondo approccio, gli Z-scores siano stati standardizzati rispetto a questa quantità.

Viene consigliato invece di utilizzare la seconda forma dell'errore standard quando si fa inferenza riguardo all'effetto casuale. In particolare è la giusta misura dell'errore da utilizzare per fare ricerca di estremi del modello.

Ecco che per testare $H_0 : \theta_i < t$ contro l'ipotesi $H_1 : \theta_i > t$, la statistica test corretta è:

$$Z_i^{(3)} = \frac{\hat{\theta}_i - t}{\mathbb{V}_{Y_i, \theta_i}(\hat{\theta}_i - \theta_i)} = \frac{\hat{\theta}_i - t}{\sigma_i\sqrt{a_i}}.$$

Se la soglia t è posta uguale a μ , il valore atteso dei θ_i , abbiamo:

$$Z_i^{(3)} = \frac{y_i - \mu}{\sqrt{\sigma_i^2/a_i}}.$$

Il relativo p-value diventa:

$$p_i^{(3)} = 1 - \Phi(Z_i^{(3)}).$$

In questo approccio, diversamente dai due precedenti, siamo quindi interessati alla ricerca delle performance di quelle unità che, pur essendo giustificate dal modello, hanno presentato valori dell'effetto casuale estremi rispetto a quelli delle altre unità; per estremi intendiamo quindi performance che stiano molto sopra (o molto sotto) la media comune μ .

3.3 L'influenza della correlazione

Osserviamo le espressioni dei p-value nei tre casi appena esposti; riscrivendo in funzione di a_i abbiamo

1. Identificazione di outlier dal modello common mean:

$$p_i^{(1)} = \Phi\left(\frac{\mu - y_i}{\sigma_i}\right)$$

2. Identificazione di outlier dal modello random effect:

$$p_i^{(2)} = \Phi\left(\sqrt{1 - a_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right)$$

3. Identificazione di estremi dal modello random effect:

$$p_i^{(3)} = \Phi\left(\sqrt{a_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right)$$

Osserviamo che se a_i è vicino a zero, allora i primi due p-value sono molto simili, mentre se a_i è vicino a 1, allora il primo p-value è simile al terzo p-value. Il valore a_i può essere interpretato come una misura della $\text{Cor}(Y_{i1}, Y_{i2})$, la correlazione tra due generiche misure fatte per la medesima unità sotto indagine, assumendo che la varianza σ_i^2 nota e costante.

Definiamo quindi la seguente misura riassuntiva della correlazione tra due misure fatte su un generico provider, 'mediato' rispetto alla dimensione:

$$r = \frac{\tau^2}{\tau^2 + \bar{\sigma}^2};$$

$\bar{\sigma}^2$ è la media pesata delle σ_i^2 . Sostanzialmente r ci mette a disposizione una misura della variabilità tra i provider confrontati, rapportata alla variabilità media totale.

Se il valore di r è vicino a zero, allora le conclusioni date dagli approcci common mean outliers e random effect outliers saranno molto simili, ovvero

evidenzieranno sostanzialmente le stesse unità come unità ‘interessanti’; infatti r vicino a zero implica che τ^2 sia anch’esso vicino a zero. Di conseguenza l’errore standard del primo e secondo approccio sono molto simili. Contemporaneamente se r si avvicina a zero, allora anche τ^2 tende a zero, ovvero la distribuzione a posteriori dei θ_i sarà stretta e centrata vicino a μ : conseguentemente, con un approccio del tipo tre, poche unità saranno identificate come *estreme* rispetto al modello random effect.

Al contrario, se r è vicino a uno, la variabilità tra i provider è alta: seguendo il secondo approccio, una variabilità più ampia sarà di conseguenza permessa, e saranno poche le unità riconosciute come outlier rispetto al modello random effect. Al contempo, le unità riconosciute come notevoli dal primo approccio tenderanno ad essere evidenziate anche dal terzo approccio.

Volendo utilizzare dei Funnel Plot per riassumere i risultati prodotti da questi procedimenti, l’effetto di r si ripercuoterà nella costruzione dei limiti di confidenza:

- Se r tende al valore zero, un Funnel Plot costruito utilizzando i limiti ricavati dal primo approccio sarà molto simile al Funnel Plot costruito utilizzando i limiti ricavati dall’approccio 2. La forma del terzo Funnel Plot sarà invece legata al valore ottenuto per la stima di τ^2 . In generale avrà limiti ampi, non identificando pochi estremi.
- Se viceversa r tende ad avvicinarsi ad uno, i grafici costruiti con i limiti previsti dal primo e dal terzo approccio si assomiglieranno molto. Il grafico prodotto dal secondo approccio presenterà invece dei limiti molto più ampi, dovuti sostanzialmente all’effetto di τ^2 che non si smorza all’aumentare della precisione come invece accade per σ_i^2 .

Nella realtà è necessario chiaramente avere una stima di r : è possibile ad esempio utilizzare la seguente stima

$$\hat{r} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \bar{\sigma}^2}.$$

dove si inserisce la stima $\hat{\tau}^2$ del parametro τ^2 .

3.4 Un esempio

Proviamo ad analizzare nuovamente i dati riportati in Tabella 1.1 e Tabella 1.2. Utilizziamo come in precedenza l’RAMR come indicatore di performace e il numero di morti attese E come indice di precisione. Per semplicità,

ricaviamo i limiti normali approssimati, ma lo stesso ragionamento potrà essere fatto anche per i limiti esatti.

Ricordiamo che per ricavare i limiti relativi all'RAMR, è sufficiente ricavarli per l'indice di precisione O/E (morti osservate fratto morti attese) e moltiplicare successivamente per la percentuale di mortalità media nel periodo in esame (nel nostro caso 2.15%). Stimiamo quindi i valori di τ^2 , $\bar{\sigma}^2$ ed r per O/E . Come target per O/E usiamo $\theta_0 = 1$. Per quanto riguarda i limiti normali per la ricerca di estremi, imponendo $\theta_0 = 1$, si esprimono in funzione del parametro di precisione E come segue:

$$\begin{aligned}
 y_p &= \theta_0 + z_p \sqrt{\frac{\sigma^2}{a}} \\
 &= \theta_0 + z_p \sqrt{\frac{\sigma^2}{\tau^2}(\tau^2 + \sigma^2)} \\
 &= \theta_0 + z_p \sqrt{\frac{1}{E\tau^2}(\tau^2 + \frac{1}{E})} \\
 &= \theta_0 + z_p \sqrt{\frac{1 + E\tau^2}{E^2\tau^2}}.
 \end{aligned} \tag{3.6}$$

Nel nostro caso otteniamo: $\hat{\tau}^2 = 0.0306$, $\bar{\sigma}^2 = 0.164$ e $\hat{r} = 0.158$ per i dati del periodo 1996-1998 e $\hat{\tau}^2 = 0.036$, $\bar{\sigma}^2 = 0.149$ e $\hat{r} = 0.196$. Per la stima di τ^2 utilizziamo il metodo già usato in precedenza, usando la stima Winsorizzata di Φ con $q = 0.1$ (ulteriori dettagli riguardo alle procedure per stime robuste possono essere trovati in [13] e in [15]). Riportiamo in Figura 3.1 i tre Funnel Plot per i dati del periodo 1996-1998 nella colonna a sinistra e per i dati in del periodo 1997-1999 nella colonna a destra.

Il valore di \hat{r} ottenuto in entrambi i casi è piuttosto piccolo; quindi le conclusioni date dai Funnel Plot relativi alla ricerca di outlier nel modello common mean e nel modello random effect sono pressochè le medesime: in entrambi i casi le uniche strutture riconosciute come outlier rispetto al modello common mean (l'ospedale di Saint Vincent nel primo periodo e l'ospedale Universitario di Brooklyn nel secondo periodo) sono identificate come outlier anche rispetto al modello random effect. Ciò è dovuto al fatto che, essendo la stima di τ^2 contenuta, in entrambi i casi si ha un aggiustamento per l'over-dispersion di piccola entità.

Molto più ampi invece sono i limiti per l'individuazione degli estremi nel modello random effect. Ciò è dovuto al fatto che, essendo i θ_i molto simili tra loro (la loro varianza τ^2 infatti è piccola in entrambi i periodo analizzati), non è possibile identificare con sicurezza quale delle realizzazioni dei θ_i giaccia sopra o sotto la media. Tant'è che in entrambi i Plot nessuna struttura giace all'infuori di entrambi i limiti tracciati.

Come anche nel capitolo precedente, la ricerca di outlier non porta a conclusioni diverse da quanto osservato nei due precedenti capitoli; gli aggiustamenti per l'over-dispersion non sono particolarmente necessari per questi due data set utilizzati negli esempi.

Da ultimo rimarchiamo nuovamente che i grafici per la ricerca degli outlier rispetto al modello common mean sono identici a quelli presentati in Figura 1.2 e in Figura 1.3, mentre quello relativo alla ricerca degli outlier nel modello random effect per il periodo 1996-1998 è equivalente a quello già riportato in Figura 2.4.

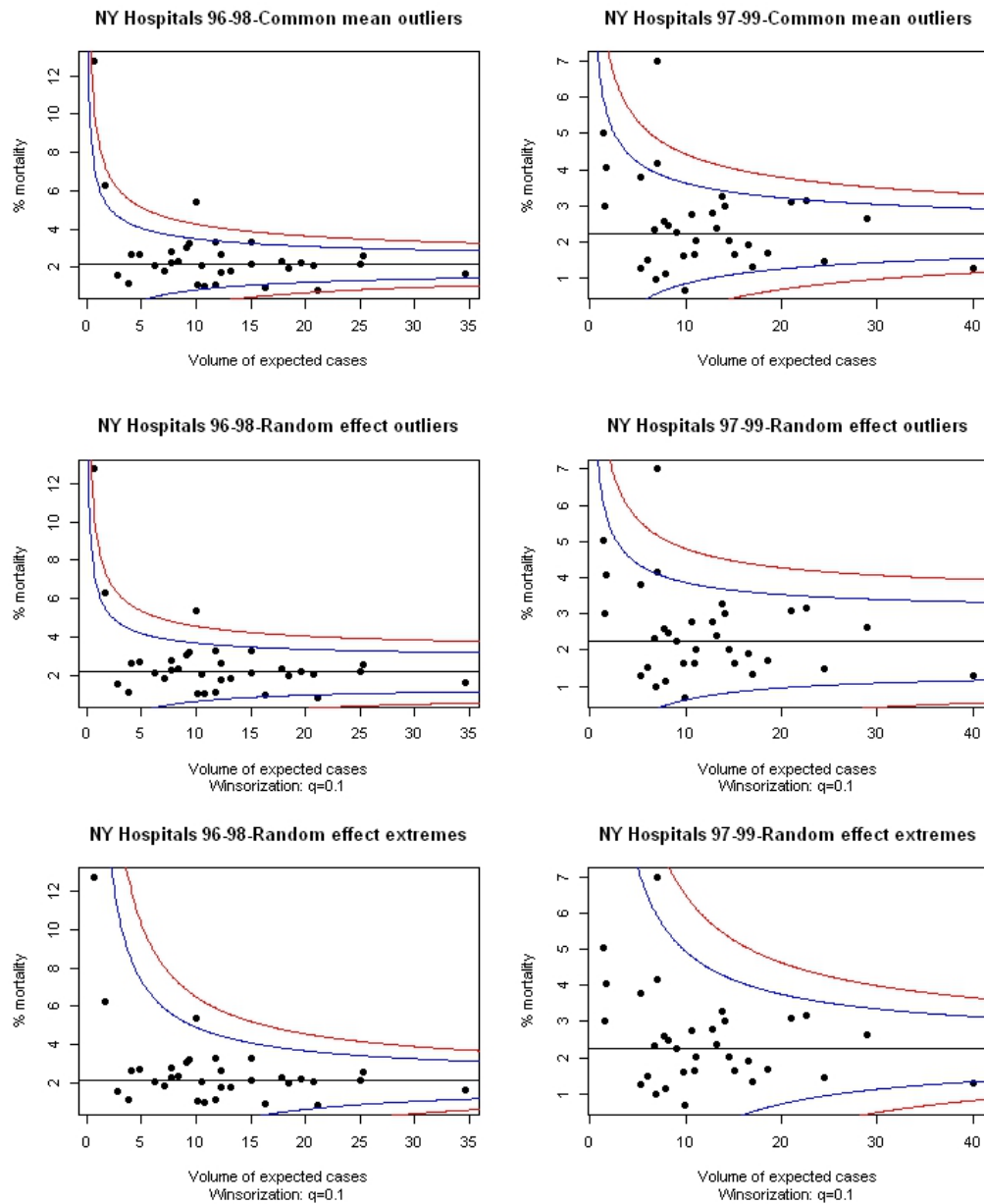


Figura 3.1: RAMR conseguente a operazioni di bypass coronarico negli ospedali dello stato di New York nel periodo 1996-1998 (a sinistra) e 1997-1999 a destra (in blu i limiti normali al 95% e in rosso i limiti normali al 99.8%).

Capitolo 4

Dati del progetto *Scompenso* di Regione Lombardia

NEI precedenti capitoli abbiamo presentato diversi esempi di utilizzo dei Funnel Plot; in particolare abbiamo sfruttato principalmente due insiemi di dati facilmente reperibili in letteratura, ovvero i dati riguardanti il progetto *CABG* dello stato di New York in due particolari periodi di tempo. Entrambi i data-set utilizzati erano piuttosto semplici in quanto già ampiamente utilizzati e trattati in primo luogo dallo stesso stato di New York, e in secondo luogo dagli autori degli articoli in cui sono stati citati (vedi [13]).

Vorremmo ora utilizzare un data set molto più ampio, riguardante il progetto di monitoraggio di Regione Lombardia circa una ampia gamma di patologie cardio-vascolari che vanno sotto il nome di scompenso cardio-circolatorio cronico. Dapprima presentiamo le informazioni contenute nel data-set a disposizione; successivamente proveremo a costruire alcuni Funnel Plot utilizzando un opportuno indice di precisione e applicando i diversi approcci studiati nei capitoli precedenti.

I dati sono stati resi disponibili all'interno della 'Ricerca Finalizzata Ministero della Salute' dell'anno 2009 **Utilization of regional health service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure**, e provengono dalla *Banca dati Assistito*, il database amministrativo di Regione Lombardia.

4.1 Presentazione del data-set

I dati riguardano numerose ospedalizzazioni¹ di pazienti affetti da scompenso cardiaco cronico (in seguito HCD: *Heart Chronic Disease*) nei principali ospedali lombardi e in alcuni ospedali extra-lombardi. Questi casi sono stati selezionati prendendo in considerazione le ospedalizzazioni occorse in una finestra temporale di dieci anni (dal 2000 al 2010). Di tutte queste ospedalizzazioni, sono state considerate solamente quelle le cui SDO (Schede di Dimissione Ospedaliera) possano essere ricondotte ad un caso classificabile come HCD. Inoltre sono selezionati come popolazione da studiare i pazienti che sono stati dimessi durante l'anno 2006 da un ospedale della regione Lombardia. Per questi pazienti sono stati quindi considerati tutti i ricoveri fino al 31 Dicembre dell'anno 2010. In particolare la prima ammissione riportata è datata 25 Luglio 2005, mentre l'ultima ammissione considerata è avvenuta il 30 Dicembre 2010.

Per ogni ospedalizzazione sono riportate numerose informazioni. Innanzi tutto un codice identificativo del paziente, totalmente anonimizzato, cui si riferisce la scheda di dimissione ospedaliera (campo `IDpt`); il codice è proprio di ogni paziente all'interno del data-set, e permette di identificare tutte le ospedalizzazioni nel periodo in esame che si riferiscono al medesimo soggetto. Per ogni paziente vi è anche l'età in anni (`age`) e il sesso (`sex`).

Tra le altre informazioni di natura generale abbiamo anche:

- `event-original`: un contatore che tiene conto degli altri *eventi* precedenti quello in questione².
- `year-discharge`: l'anno di dimissione.
- `DaystoDead`: il numero di giorni tra la dimissione del paziente e la morte dello stesso. Se la morte è avvenuta oltre la fine del periodo di interesse (oltre il Dicembre 2010) il valore è da considerare censurato a destra.
- `Date-ADMISSION`: data di ricovero. Coincide colla data della prima ammissione se l'evento ha coinvolto anche trasferimenti tra diversi ospedali.

¹Il data set completo contiene informazioni riguardanti 72992 schede di dimissione, per un totale di 28095 pazienti distinti.

²Notiamo che il primo ricovero di un paziente può non necessariamente coincidere con `event-original=1`, nel caso ad esempio il paziente sia già stato ricoverato in un periodo precedente quello di studio per problemi analoghi.

- **Date-DISCHARGE**: data di dimissione. Se l'evento ha coinvolto anche trasferimenti tra diversi ospedali, coincide con la data di dimissione dall'ultimo ospedale.
- **Date-DEATH**: data di morte del paziente.
- **Date-1ev**: data della prima ammissione per ogni paziente.
- **Date-from-1ev**: numero di giorni dal primo evento che ha coinvolto il paziente in questione.

Sono presenti anche informazioni riguardanti eventuali altri eventi sensibili avvenuti precedentemente al ricovero dal paziente.

- **SHOCK**: indica se il paziente ha subito uno evento di shock cardiogeno in almeno una delle ammissioni avvenute durante l'evento considerato.
- **CABG**: indica se il paziente ha subito in passato un'operazione di bypass coronarico.
- **PTCA**: indica se il paziente ha subito in passato un'operazione di angioplastica.
- **ICD**: indica se il paziente ha avuto bisogno di un *implantable cardiac defibrillator*.
- **STENT**: indica se il paziente ha avuto bisogno di uno Stent cardiaco.

Per ogni paziente abbiamo anche una serie di indicatori binari di comorbidità³. Ognuno di essi indica se nell'evento considerato (o in uno dei precedenti) il paziente è stato affetto anche da vari altri disturbi. Per ogni indice di comorbidità è riportato il nome del disturbo (ad esempio *arrhythmia* per l'aritmia cardiaca) e l'indice utilizzato per determinarne la presenza; in particolare sono stati utilizzati due possibili indici, ovvero *Romano* ed *Elixhauser*, dai nomi dei relativi sviluppatori. Per esempio la variabile *arrhythmia-elixhauser* vale 1 se nell'evento considerato o in uno dei precedenti ricoveri il paziente ha sofferto di aritmia cardiaca, in accordo con l'indice di Elixhauser. Le comorbidità riportate nel data-set a disposizione riguardano i seguenti disturbi: tumore, cancro allo stato metastatico, demenza, psicosi, insufficienza renale, insufficienza cardiaca, ipertensione, anemia,

³Nel linguaggio medico, per comorbidità (o comorbilità) si intende la presenza per lo stesso paziente di più disturbi di origine diversa.

coagulopatia, diabete, disturbi epatici, perdita di peso, insufficienza polmonare, difficoltà circolatoria a livello polmonare, difficoltà circolatoria a livello periferico, abuso di alcool, disturbi motori quali l'emiplegia, AIDS.

Le seguenti informazioni si riferiscono invece agli ospedali coinvolti:

- **ID-hosp**: un codice identificativo numerico dell'ospedale di ammissione del paziente. Se l'ospedalizzazione ha coinvolto anche trasferimenti tra ospedali, si riferisce alla prima ammissione.
- **reg-hosp**: un codice identificativo della regione cui l'ospedale della prima ammissione appartiene. Osserviamo che, nonostante ci si stia occupando delle schede di dimissione compilate in ospedali lombardi, la prima ammissione del paziente potrebbe essere avvenuta in ospedali extra-lombardi.
- **hosp-lomb**: vale 1 se l'ospedale di prima ammissione appartiene alla regione Lombardia. Osserviamo che le ammissioni avvenute fuori dalla regione Lombardia sono 1.687.
- **dead-intraH**: vale 1 se il paziente è morto durante il periodo di ricovero. In questo caso la data di morte coincide con la data di dimissione relativa al ricovero.
- **n-trasf**: indica il numero di trasferimenti subiti dal paziente durante l'evento considerato.

Prima di proseguire facciamo alcune osservazioni. Prima di tutto, nonostante lo studio abbia riguardato pazienti dimessi da ospedali lombardi, vi sono tuttavia ben 1522 schede riguardanti pazienti ricoverati in ospedali non lombardi e che non hanno subito alcun trasferimento. Si tratta di pazienti che avevano già subito un precedente ricovero con almeno una dimissione da un ospedale lombardo. Le schede di dimissione relative a questi ricoveri in ospedali non lombardi sono state poi raggruppate in quanto i ricoveri in questione si sono avvenuti in tempi successivi. In ogni caso non vi sono schede riguardanti pazienti ricoverati per tutto il periodo di degenza in un ospedale non lombardo e che contemporaneamente fossero al primo ricovero per problemi di scompenso cardio-circolatorio.

Osserviamo anche che per alcune schede vi sono errori di compilazione legati alle date di ricovero e di dimissione (date di ammissione successive a date di dimissione o a date di morte, e altri errori simili dovuti probabilmente alla compilazione errata da parte dell'operatore). Bisognerà innanzi tutto eliminare le righe relative a queste schede con date scorrette. In particolare

vi sono 295 schede scorrette sotto questo punto di vista e che non considereremo. Rimangono quindi 72697 schede di dimissione (corrispondenti a 27995 pazienti differenti).

Eliminiamo poi quelle schede aventi `n-trasf` maggiore di zero, cioè eliminiamo le schede riguardanti ricoveri in cui il paziente abbia subito dei trasferimenti tra ospedali. Questo perchè successivamente vorremo poter associare ogni paziente ad un singolo ospedale; se il numero di trasferimenti è maggiore di zero, all'interno del particolare ricovero non è possibile risalire agli ospedali visitati dal paziente. Il numero di pazienti si riduce ulteriormente a 26650, di cui 13236 sono maschi e 13414 sono femmine. Nelle Figure 4.1, 4.2 ed 4.3 riportiamo, rispettivamente, i box plot e gli istogrammi per le età di tutti i pazienti, dei soli pazienti maschi e delle sole pazienti femmine. Possiamo vedere che le popolazioni dei maschi e delle femmine sono confrontabili per distribuzione delle età.

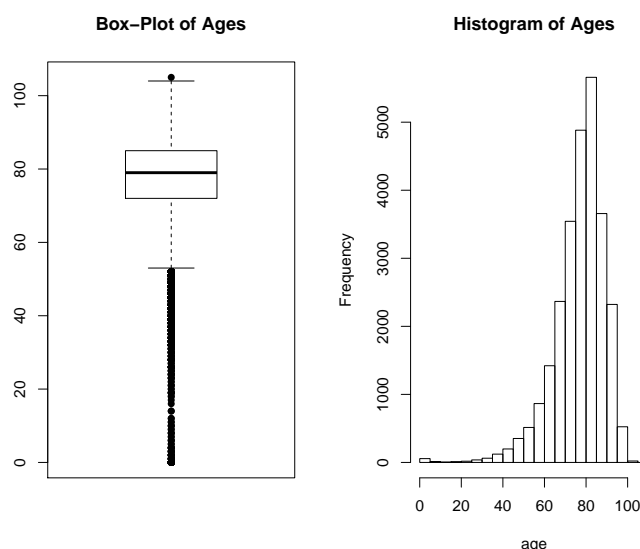


Figura 4.1: Box-Plot e istogramma delle età dei pazienti.

4.2 Funnel Plots

Utilizziamo i dati appena presentati per realizzare, con le tecniche descritte precedentemente, alcuni Funnel Plot. Come indice di precisione utilizzeremo il volume di pazienti avuti in degenza per ogni ospedale. Un paziente che sia stato ricoverato più volte nello stesso ospedale in tempi successivi verrà

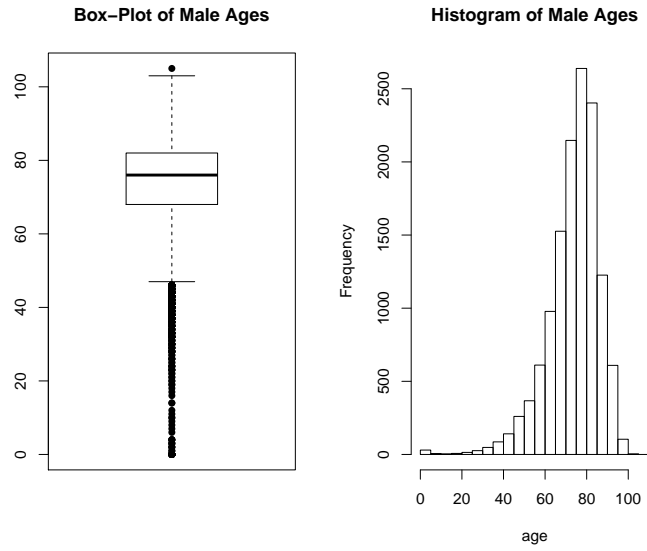


Figura 4.2: Box-Plot e istogramma delle età dei pazienti maschi.

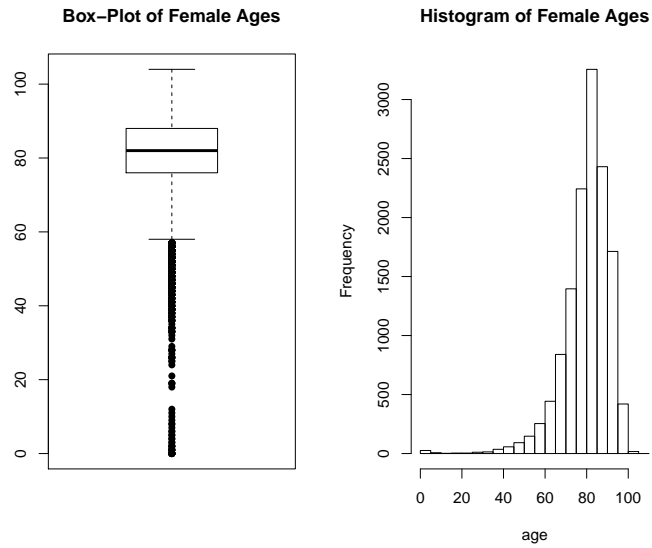


Figura 4.3: Box-Plot e istogramma delle età delle pazienti femmine.

contato una sola volta. Non siamo quindi interessati alle singole ospedalizzazioni dei pazienti, ma al singolo paziente che sia stato ricoverato in un ospedale (eventualmente più volte). Notiamo però che un singolo paziente può essere contato in più ospedali. Come risposta avremo invece la mortalità (espressa in percentuale), calcolata in relazione ai pazienti morti nell'ospedale in questione. Come valore di target avremo la mortalità media (pesata per le numerosità). Come già sottolineato, ci concentriamo solo su quei pazienti ricoverati e dimessi dal medesimo ospedale, cioè solo su quelle schede di dimissione in cui `n-trasf` vale zero. Di conseguenza il numero ospedali da esaminare si riduce a 497 sui 510 originali.

Per evitare di creare nuvole troppo concentrate, e soprattutto per indagare gli ospedali che abbiano avuto a che fare con un numero sufficientemente grande di pazienti, ci restringiamo a quelle strutture che hanno trattato non meno di 20 individui nel periodo 2006-2010. In tal maniera riduciamo il numero di ospedali da rappresentare nei grafici da 497 a 163.

4.2.1 Funnel Plot senza correzione per l'over-dispersion

Seguendo le indicazioni riportate nel Capitolo 1 alla sezione riguardante le proporzioni, possiamo ricavare i limiti di controllo normali, la cui espressione era:

$$y_p = \theta_0 + z_p \sqrt{\frac{\theta_0(1 - \theta_0)}{n}}.$$

I limiti riportati, come nel Capitolo 1, sono al 95% e 99.8%.

In questo caso θ_0 è circa il 15%. Riportiamo in Figura 4.4 il plot costruito. È immediato vedere che 18 ospedali hanno registrato una mortalità oltre il limite di controllo superiore. In particolare sono gli ospedali identificati dai seguenti codici: 36, 37, 42, 58, 64, 69, 73, 80, 110, 111, 141, 142, 164, 165, 178, 197, 198 e 204. Allo stesso modo si possono identificare i 19 ospedali identificati come particolarmente virtuosi, cioè quelli che riportano una mortalità più bassa del limite di controllo inferiore: 44, 45, 91, 96, 113, 119, 131, 132, 134, 148, 153, 160, 169, 213, 218, 220, 222, 228 e 231. Come già osservato, vi sono ospedali che presentano mortalità superiori ad alcuni ospedali classificati come fuori controllo, ma l'effetto della precisione ci permette di tener conto di quanto affidabili siano queste mortalità, alla luce del modello utilizzato per rappresentare la variabilità.

Vi sono poi alcuni ospedali confinati nella zona di *warning* tra i limiti più stretti e quelli più larghi: in particolare ve ne sono 12 nella fascia di *warning* superiore (identificati dai codici 38, 51, 59, 67, 71, 72, 75, 104, 121, 137, 205

e 216), e ben 35 ospedali nella fascia inferiore. I rimanenti 79 ospedali sono considerati sotto controllo da questo Funnel Plot.

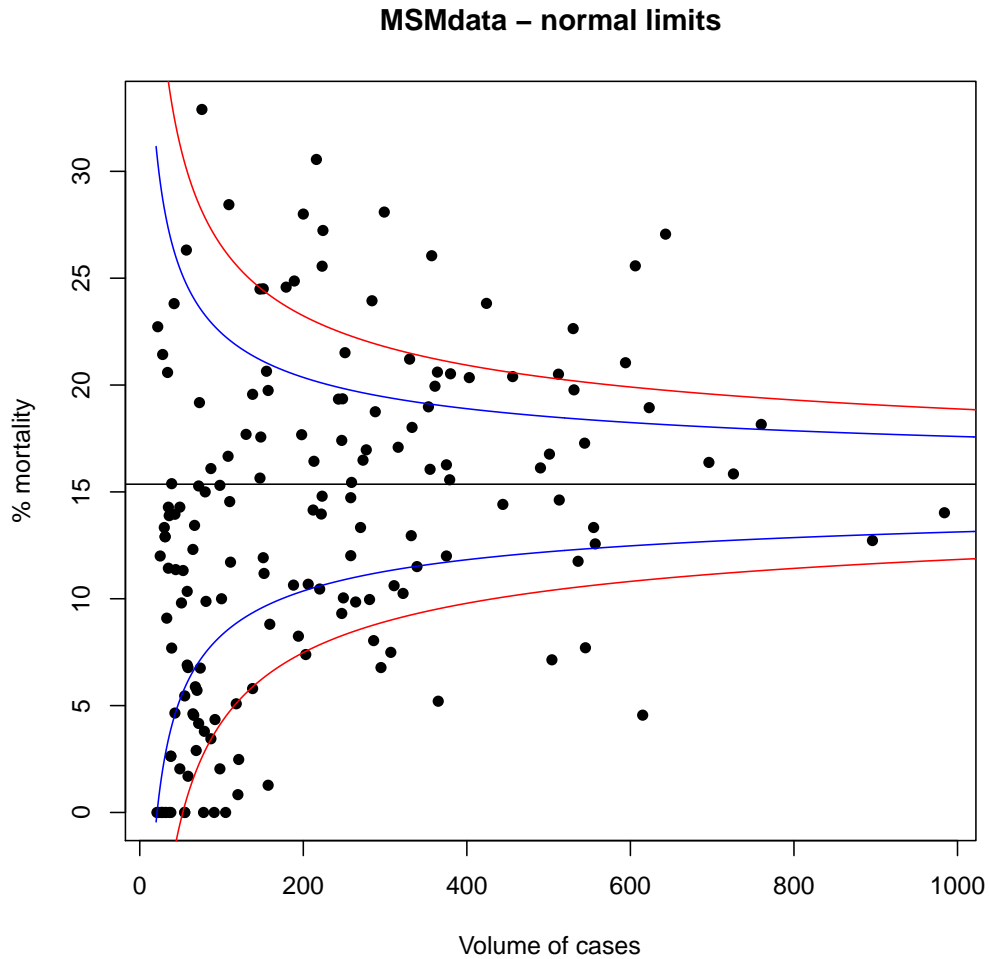


Figura 4.4: Funnel Plot per i dati in esame, senza correzioni per l'over-dispersion.

4.2.2 Correzioni per l'over-dispersion

Usiamo qui le due tipologie di correzione (moltiplicativa e additiva) come introdotte nel Capitolo 2.

Correzione moltiplicativa

Per prima cosa è necessario calcolare la stima di ϕ utilizzando la formula (2.2). Calcoliamo quindi dapprima gli Z-scores che in questo caso avranno la seguente espressione:

$$Z_i = \frac{y_i - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}.$$

In Figura 4.5 è rappresentato il grafico dei quantili normali degli Z scores. Possiamo vedere che i valori dei residui sono molto grandi; potrebbe essere quindi necessario utilizzare una stima *Winsorised* per ϕ ; infatti la stima diretta $\hat{\phi}$ senza Winsorization è molto grande, pari a 7.21.

Proviamo ad utilizzare due diversi valori di q per calcolare la stima *Winsorised*, ovvero $q = 0.1$ e $q = 0.2$. Nel primo caso otteniamo $\hat{\phi} = 4.63$, nel secondo caso abbiamo $\hat{\phi} = 2.9575$. In Figura 4.6 abbiamo i due grafici corretti con il termine di *inflation* ϕ .

Nel grafico a destra si può notare che i limiti di controllo sono stati molto ampliati. Ora solo due strutture sono indicate come fuori controllo, superando il limite superiore. In particolare si tratta degli ospedali indicati dal codice 164 e 197. Al contempo solo una struttura è indicata come particolarmente virtuosa, riportando una mortalità inferiore al limite più basso, ovvero l'ospedale 222. Nella fascia di warning superiore abbiamo 9 ospedali (36, 37, 58, 64, 69, 80, 142, 165 e 178) mentre in quella inferiore ne abbiamo 6 (gli ospedali 44, 113, 169, 213, 228 e 231).

Nel grafico di sinistra vengono indicate 5 strutture sopra il limite superiore e due sotto quello inferiore. Per quanto riguarda le strutture con mortalità troppo alta, vengono ora segnalate anche gli ospedali 64, 142 e 178. All'ospedale numero 222 si aggiunge invece il numero 113. Nella fascia di warning superiore abbiamo ora 11 ospedali (le strutture numero 36, 37, 58, 69, 73, 80, 110, 111, 141, 165 e 204). Abbiamo ancora 12 ospedali nella fascia di warning inferiore (gli ospedali 44, 45, 132, 134, 148, 160, 169, 213, 218, 220, 228 e 231).

Osserviamo che entrambi i due grafici concordano nell'indicare alcune strutture come fuori controllo, ovvero quelle più estreme. Forse è preferibile la stima di *winsorized* ϕ con $q = 0.2$. Questo valore permette di contenere maggiormente il valore stimato di ϕ . È comunque chiaro da quanto visto che le strutture 164 e 197 hanno un comportamento anomalo. La struttura 22 mantiene secondo entrambi i Funnel Plot una performance particolarmente virtuosa.

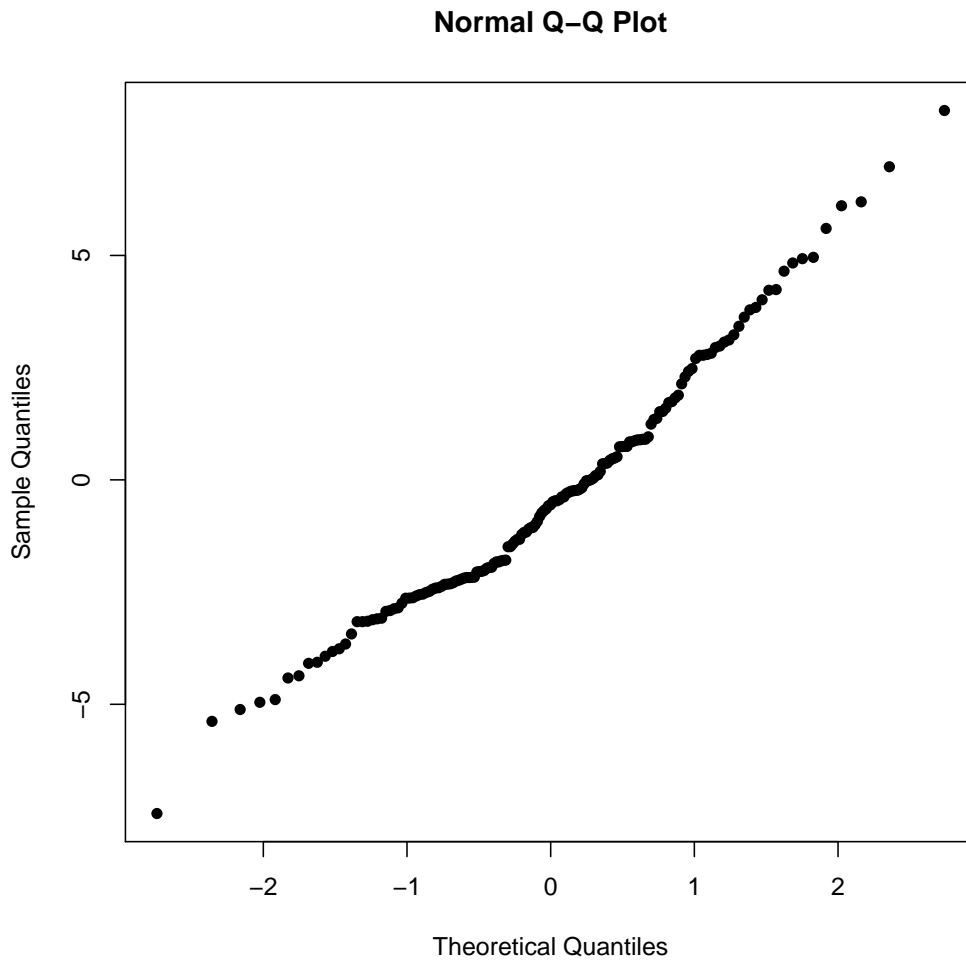


Figura 4.5: Grafico dei quantili normali per i residui di Pearson per i dati in esame.

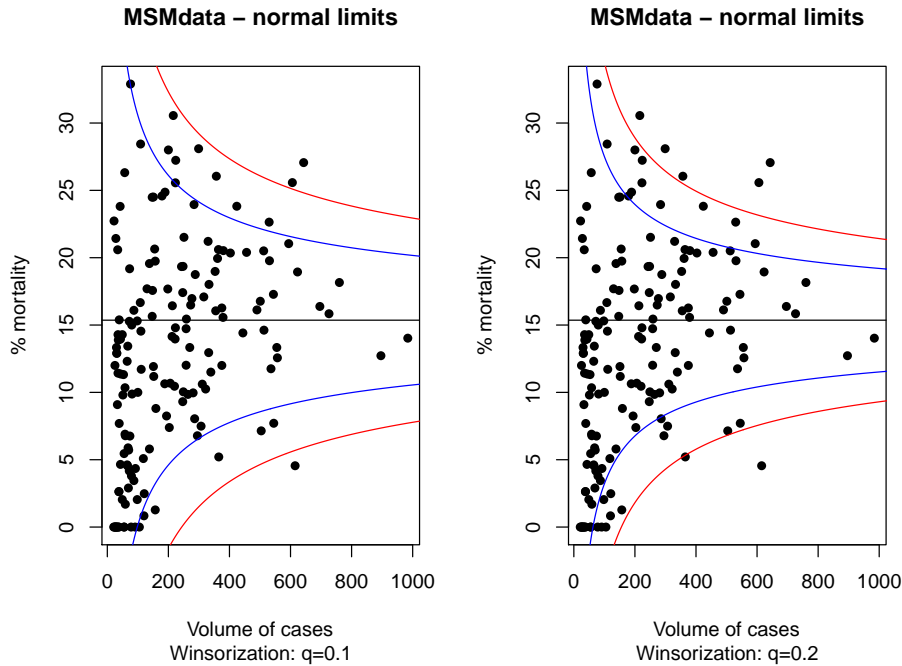


Figura 4.6: Funnel Plots con correzione moltiplicativa per i dati in esame.

Correzione additiva

Possiamo utilizzare le formule (2.2), (2.4) e (2.5), come presentate nel Capitolo 2 per stimare la componente casuale additiva τ^2 . Ancora una volta per la stima di ϕ usiamo i due procedimenti di Winsorization precedenti, con $q = 0.1$ e $q = 0.2$. I valori di $\hat{\tau}^2$ che si ottengono sono 21.6068 per $q = 0.1$ e 11.66875 per $q = 0.2$. In Figura 4.7 sono riportati i due grafici corretti con la componente additiva.

Osserviamo subito che nel grafico di sinistra non vengono identificate strutture oltre i limiti superiore e inferiore di controllo. Questo perchè la stima di τ^2 è molto elevata, determinando una componente consistente della variabilità che non viene abbattuta all'aumentare della precisione. Nella fascia di warning superiore abbiamo 9 ospedali, mentre nella fascia inferiore ne abbiamo 15. In particolare nella fascia superiore ritroviamo ancora gli ospedali 23, 142, 164, 178 e 197 già segnalati in precedenza. Nella fascia inferiore ritroviamo l'ospedale 222, e altre strutture già segnalate.

Nel grafico di destra i limiti sono lievemente più stretti e permettono di indicare 4 strutture oltre il limite al 99.8% (ancora il 142, il 164 e il 178 cui si raggiunge l'80, già evidenziato dal grafico senza correzioni). Vengono indicati anche gli ospedali 44 e 213 sotto il limite inferiore. Osserviamo però

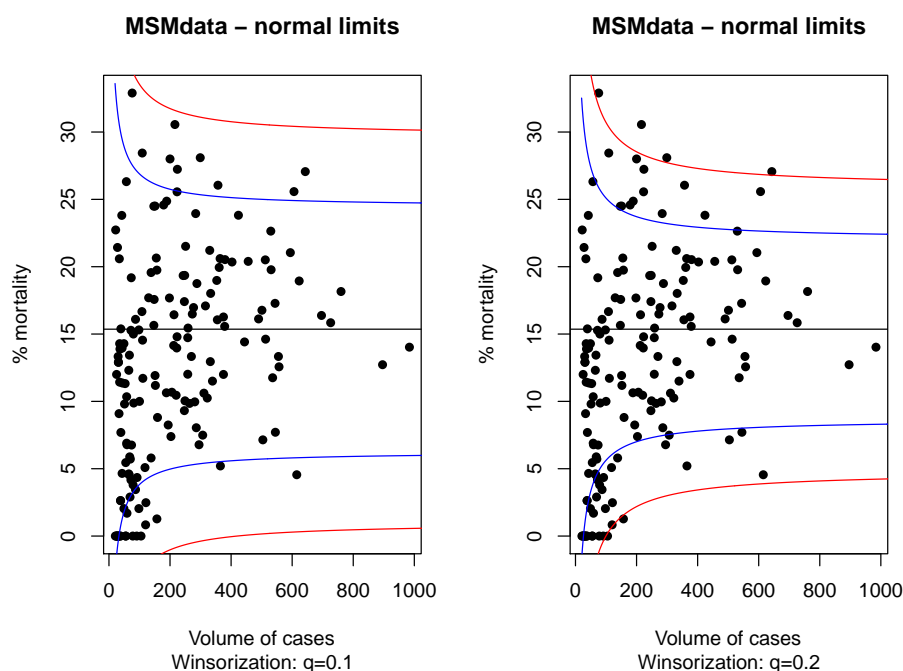


Figura 4.7: Funnel Plots con correzione additiva per i dati in esame.

che si trovano pressochè sul limite. Nelle fasce di warning cadono ancora numerose strutture già indicate in precedenza.

Quale delle due scelte di winsorization sia preferibile ancora una volta non è chiaro. Probabilmente è meglio la seconda scelta, che permette di mettere in evidenza più strutture; questo nell'ottica di optare per una linea meno conservativa.

4.3 Utilizzo del modello *random effect*

Sfruttiamo ora le tecniche introdotte nel Capitolo 3, dove abbiamo mostrato come sfruttare un modello *random effect* per rappresentare la variabilità dell'indicatore Y . In un modello di questo tipo ogni struttura è caratterizzata da una media θ_i proveniente da una distribuzione normale, con media comune μ e una varianza τ^2 . Il termine τ^2 rappresenta quindi la variabilità *tra* le diverse strutture, e viene di conseguenza considerata nella variabilità complessiva dell'indicatore di performance Y .

Sempre nel Capitolo 3 abbiamo rimarcato la differenza di significato tra outlier ed estremi: quando ricerchiamo gli outlier siamo interessati a quegli ospedali che hanno presentato nel periodo in esame un indice di performance

Tabella 4.1: Valori delle stime di ϕ , τ^2 e r .

q	$\hat{\phi}$	$\hat{\tau}^2$	\hat{r}
0.00	7.21	36.94	0.72
0.10	4.63	21.61	0.60
0.20	2.96	11.67	0.45
0.30	1.96	5.75	0.29

‘fuori’ dai limiti del modello; se siamo interessati invece alla ricerca degli estremi, la nostra attenzione è rivolta a quelle strutture cliniche la cui media θ_i si allontana maggiormente dalla media comune.

Abbiamo anche visto tre possibili metodi per identificare ospedali fuori controllo, sottolineando il ruolo centrale rivestito dalla scelta della stima della variabilità utilizzata. Per la ricerca di outliers dal modello common mean viene utilizzato come stima della variabilità la stessa σ_i^2 già usata in precedenza, mentre per la ricerca di outliers dal modello a effetto casuale si somma a σ_i^2 la stima $\hat{\tau}^2$ (in maniera equivalente, facciamo notare ancora una volta, all’approccio di correzione additiva). Prestiamo un momento di attenzione invece alla variabilità da utilizzare per la ricerca degli estremi: l’errore standard da utilizzare vale, per quanto detto nel Capitolo 3 $\sqrt{\sigma_i^2/a_i}$, per cui si ha:

$$\begin{aligned}
 \frac{\sigma_i^2}{a_i} &= \sigma_i^2 \frac{\sigma_i^2 + \tau^2}{\tau^2} \\
 &= \frac{\theta(1-\theta)}{n\tau^2} \left(\frac{\theta(1-\theta)}{n} + \tau^2 \right) \\
 &= \frac{\theta(1-\theta)}{n^2\tau^2} (\theta(1-\theta) + n\tau^2)
 \end{aligned} \tag{4.1}$$

La quantità τ^2 viene poi stimata usando la formula (3.4). Osserviamo che è necessario una preventiva stima della quantità ϕ per poter stimare l’effetto casuale. Si ripresenta quindi ancora il problema riguardante quale stima utilizzare, ovvero quale (eventuale) Winsorization applicare. Decidiamo di applicare quattro diversi livelli: nessuna Winsorization, $q = 0.1$, $q = 0.2$ e $q = 0.3$.

In Tabella 4.1 sono riportati i valori delle diverse stime di ϕ , τ^2 ed r nel caso in esame. Per la stima di r è necessario stimare la quantità $\bar{\sigma}^2$ che rappresenta la ‘media pesata’ delle σ_i^2 . Per i dati utilizzati vale $\bar{\sigma}^2 = 14.39628$.

Nelle figure 4.8, 4.9, 4.10 e 4.11 riportiamo i Funnel Plot costruiti per i nostri dati utilizzando l'approccio con il *random effect model*. In ogni tripletta di grafici abbiamo a sinistra l'approccio common mean outliers, che abbiamo già osservato essere equivalente al Funnel Plot senza correzioni. In centro il grafico che segue l'approccio 2 (random effect outliers) e a destra il grafico basato sulla ricerca degli estremi nel modello a effetto casuale.

Una visione di insieme ci permette innanzi tutto di apprezzare l'effetto della stima del coefficiente r sulla forma complessiva dei grafici. Scorrendo le figure dalla prima alla quarta è possibile vedere un restringimento dei limiti nei grafici centrali e un conseguente allargamento degli stessi nei grafici di destra. Abbiamo già sottolineato questo effetto nel Capitolo 3. Ricordiamo che un valore basso di r consiste in un valore basso di τ^2 . Di conseguenza ciò porta i limiti del grafico centrale ad essere più stretti, in quanto l'entità dell'effetto casuale diminuisce. Al contempo i limiti del grafico di destra si allargano; questo perchè, diminuendo la variabilità dell'effetto casuale θ_i , le relative realizzazioni sono più vicine alla media comune μ e quindi sarà più difficile classificare una struttura come estrema. Vicersa, all'aumentare di r aumenta anche τ^2 , allargando i limiti del grafico centrale (viene consentita una maggiore variabilità) e restringendo contemporaneamente i limiti del grafico di destra (le strutture possono presentare realizzazioni dell'effetto casuale più ampie).

Cominciando dalla Figura 4.8 possiamo notare l'entità considerevole della stima di $\hat{\tau}^2$. In tal caso i limiti corretti con l'effetto casuale nel grafico centrale sono talmente larghi che i due limiti di controllo superiore e inferiore non compaiono nemmeno nel grafico. In questo grafico nessun ospedale è indicato come outlier e pochissimi altri cadono nella fascia di warning. Riguardo al grafico per la ricerca degli estremi, notiamo che è molto simile al grafico di sinistra, identificando sostanzialmente le stesse strutture (vedi il precedente paragrafo riguardo il Funnel Plot senza correzioni).

Un discorso analogo vale per i grafici riportati in Figura 4.9. La stima $\hat{\tau}^2$ è diminuita rispetto al primo caso, ma rimane piuttosto consistente. Per quanto riguarda il grafico centrale, esso è equivalente al grafico di sinistra riportato precedentemente in Figura 4.7. Il grafico di destra non differisce sostanzialmente dal corrispettivo di Figura 4.8.

In Figura 4.10 la situazione rimane ancora simile. Il grafico di mezzo è uguale al grafico di destra riportato in Figura 4.7. Il grafico di destra è lievemente più largo.

Nell'ultima figura, la 4.11 la situazione varia poco sensibilmente e le strutture già identificate precedentemente vengono in linea di massima confermate come oltre i limiti.

Quale delle precedenti Winsorizzazioni sia quella corretta non è noto a priori. Ancora una volta facciamo osservare che l'utilizzo di questa procedura di stima serve a rendere il valore $\hat{\phi}$ robusto rispetto all'influenza di strutture le cui performance sono fuori controllo. Bisogna quindi pensare al valore di q come ad un parametro di *tuning*, da regolare per mettere più o meno in risalto le performance di diverse strutture.

Osservando i grafici, la stima diretta di ϕ è probabilmente poco esatta. Insomma, un piccolo livello di *Winsorization* è sempre consigliabile. Valori tra 0.1 e 0.2 sembrano abbastanza buoni. Valori troppo elevati potrebbero probabilmente sottostimare ϕ e di conseguenza l'effetto casuale τ^2 .

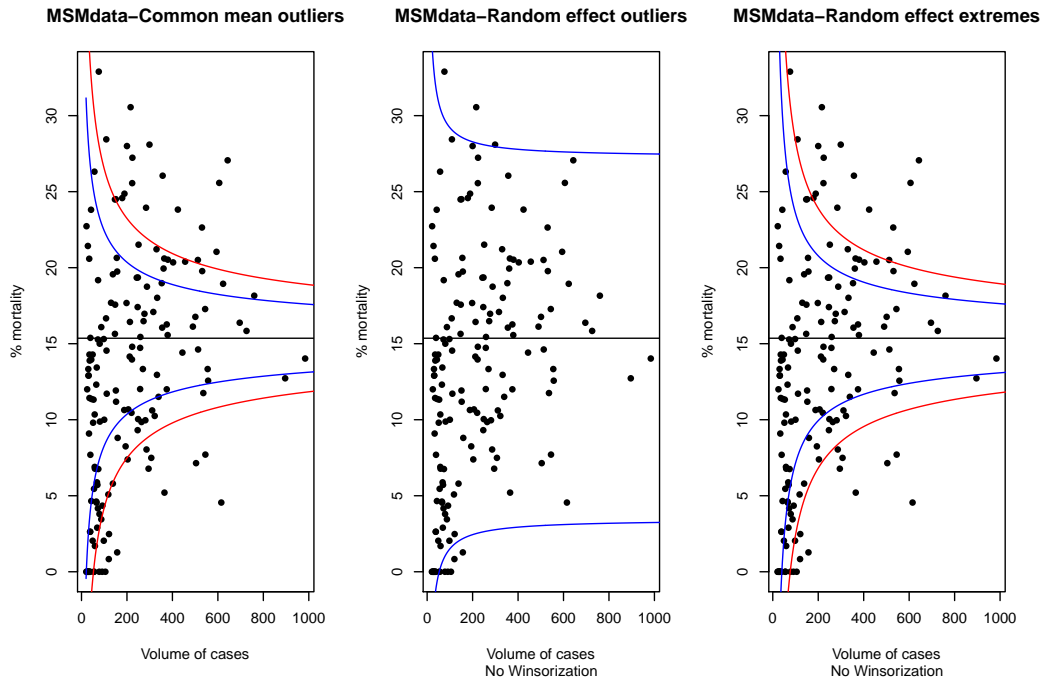


Figura 4.8: Funnel Plots costruiti utilizzando un modello ad effetto casuale ($q = 0$).

4.4 Un'osservazione prima di proseguire

I Funnel Plot costruiti in questo capitolo non utilizzano un modello che tenga conto delle caratteristiche dei pazienti trattati dai diversi ospedali. Abbiamo infatti utilizzato come indice di precisione la numerosità dei singoli pazienti trattati da ogni ospedale e come indice di performance la mortalità tra questi pazienti.

Nel Capitolo 1 abbiamo visto come è possibile usare l'SMR come indicatore, utilizzando come relativo indice di precisione l'expectation E . Questa quantità deve essere predetta attraverso un opportuno modello statistico che tenga conto possibilmente delle caratteristiche di ogni paziente che viene trattato dalla struttura.

Nei prossimi due capitoli introdurremo quindi gli elementi basilari dei Modelli Lineari Generalizzati (GLM) e dei Modelli Lineari Generalizzati ad Effetti Misti (GLMM). Nel Capitolo 7 utilizzeremo un opportuno modello GLM per costruire i Funnel Plot dei dati relativi all'HCD per la Regione Lombardia, facendo opportune riflessioni e confronti.

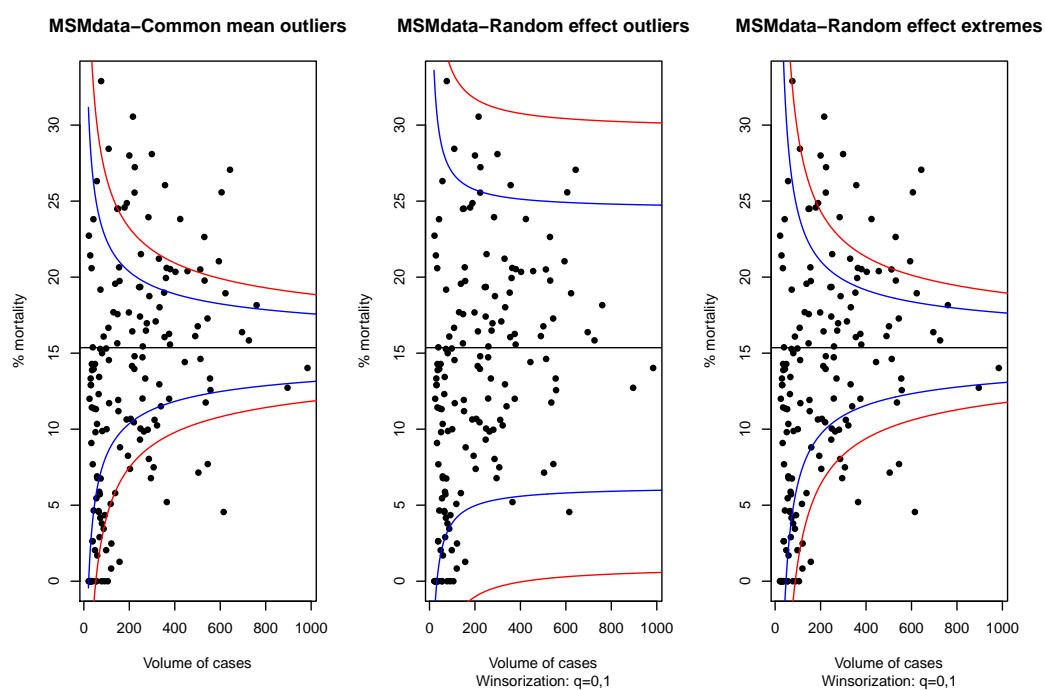


Figura 4.9: Funnel Plots costruiti utilizzando un modello ad effetto casuale ($q = 0.1$).

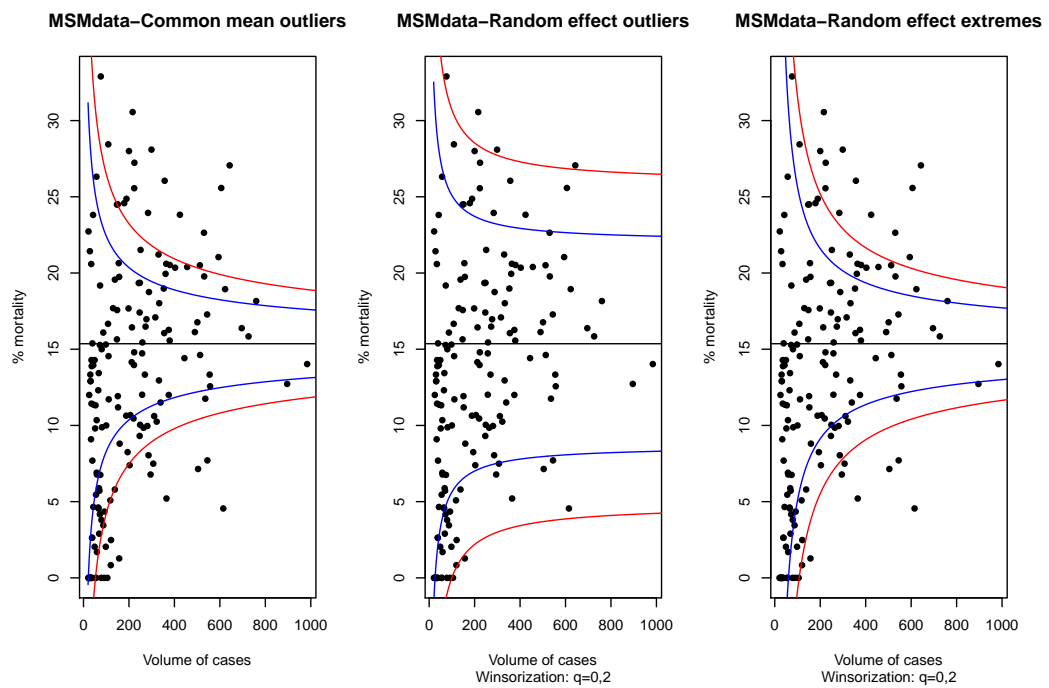


Figura 4.10: Funnel Plots costruiti utilizzando un modello ad effetto casuale ($q = 0.2$).

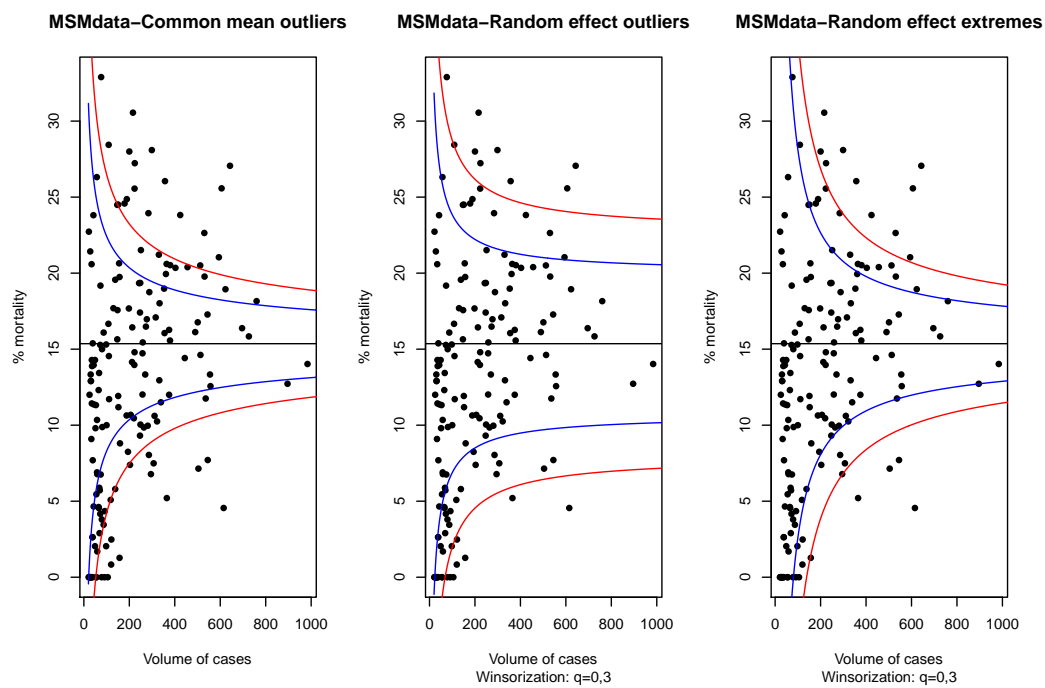


Figura 4.11: Funnel Plots costruiti utilizzando un modello ad effetto casuale ($q = 0.3$).

Capitolo 5

GLM e Regressione Logistica

PRESENTIAMO in questo capitolo i principali aspetti dei GLM (*Generalized Linear Models*) al fine di spiegare una variabile di tipo binario¹, in relazione ad altre variabili, o predittori. In casi del genere, le tecniche classiche di regressione lineare perdono utilità in quanto non si ha a che fare con risposte continue. La nostra attenzione si sposterà di conseguenza sulla modellazione di opportune funzioni della media. Per questo entrano in gioco i modelli lineari generalizzati, con lo scopo di estendere le normali tecniche di regressione. Successivamente ci concentreremo in particolare su un modello tra i più importanti, ovvero la *Regressione Logistica*.

Non ci occupiamo degli aspetti numerici riguardanti la soluzione delle equazioni di massima verosimiglianza necessarie per il calcolo delle stime dei coefficienti dei modelli considerati, sottolineando il fatto che un qualsiasi software statistico (quale ad esempio **R**, **SAS** o **Minitab**) sono in grado di fornire le stime desiderate nonché il calcolo di molte altre quantità di interesse. Per il lettore interessato rimandiamo ad [1], capitoli 4 e 5, cui si rifà in linea di massima il materiale contenuto di questo capitolo.

5.1 Principali elementi di un GLM

Prima di tutto, comprendiamo in cosa consiste un modello lineare generalizzato indicando quali elementi servono per specificarlo completamente.

- Una componente casuale (*random component*): consiste di una variabile risposta Y con N osservazioni indipendenti (y_1, \dots, y_N) provenienti da una distribuzione appartenente alla famiglia esponenziale naturale. Ciò significa che ogni osservazione proviene da una densità di

¹Ad esempio un indicatore di sopravvivenza di un paziente in un determinato ricovero.

probabilità della forma:

$$f(y_i; \theta_i) = a(\theta_i)b(y_i) \exp[y_i Q(\theta_i)]. \quad (5.1)$$

Casi particolari sono la distribuzione di Poisson (per *count data*) e soprattutto la distribuzione Binomiale (per dati binari); quest'ultima in particolare verrà utilizzata in seguito. Il valore del parametro θ_i può variare al variare dell'indice in relazione alle covariate che caratterizzano l'osservazione in esame. Il termine $Q(\theta_i)$ è detto parametro naturale.

- Una componente sistematica (*systematic component*): ha lo scopo di mettere in relazione un vettore $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)$ con le covariate (o predittori). Indicando con x_{ij} il valore del predittore j per il soggetto i , abbiamo:

$$\eta_i = \sum_{j=1}^N \beta_j x_{ij}, \quad i = 1, \dots, N. \quad (5.2)$$

In notazione matriciale:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \quad (5.3)$$

Generalmente la covariata x_{i1} è identicamente 1 per ogni soggetto, introducendo nel modello un'intercetta. La matrice \mathbf{X} è chiamata *model matrix*, ed è l'equivalente della *design matrix* per la regressione lineare classica.

- Una *link function*: è una funzione g che mette in relazione le due componenti casuale e sistematica. Se definiamo $\mu_i = \mathbb{E}(Y_i)$, la link function agisce mettendo in relazione η_i e μ_i come segue:

$$g(\mu_i) = \eta_i = \sum_{j=1}^N \beta_j x_{ij}, \quad i = 1, \dots, N. \quad (5.4)$$

Per esempio, $g(\mu) = \mu$ è detta *identity link*; è la funzione di link normalmente utilizzata nella regressione ordinaria con risposte gaussiane. Un altro esempio è la *canonical link*, che trasforma la media nel parametro naturale: $g(\mu_i) = Q(\theta_i)$.

In Tabella 5.1 è riportato un breve specchietto riassuntivo delle più consuete combinazioni di componente casuale, sistematica e link function. Ricordiamo che la trasformazione Logit è definita nel modo seguente:

Tabella 5.1: Tipi di modelli lineari generalizzati (vedi [1])

Componente Casuale		Componente Sistematica	
	Link		Modello
Normale	Identità	Continua	Regressione
Normale	Identità	Categorica	ANOVA
Normale	Identità	Mista	ANCOVA
Binomiale	Logit	Mista	Regressione Logistica
Poisson	Log	Mista	Log-Lineare

$$\text{logit}(x) = \log \frac{x}{1-x}.$$

Indicate quindi le principali componenti del modello lineare generalizzato, ci concentriamo sulla categoria di modelli più adatta alle risposte di nostro interesse, ovvero i modelli per dati binari.

5.2 GLM per risposte binarie

Concentriamoci sulle risposte di tipo binario e presentiamo un breve campionario dei modelli utili per analizzarli. Indichiamo con Y una variabile aleatoria di risposta che possa assumere due valori, indicati per convenzione con 0 e 1. Assumendo che ogni risposta sia distribuita come una variabile aleatoria di Bernoulli abbiamo:

$$\begin{aligned} f(y; \pi) &= \pi^y (1 - \pi)^{1-y} \\ &= (1 - \pi) [\pi / (1 - \pi)]^y \\ &= (1 - \pi) \exp\left(y \log \frac{\pi}{1 - \pi}\right) \end{aligned}$$

La verosimiglianza è così espressa nella forma indicata dalla (5.1), dove $a(\pi) = 1 - \pi$, $b(y) = 1$ e $Q(\pi) = \log \frac{\pi}{1 - \pi}$, ovvero il Logit di π . Ricordiamo che il parametro π ha il significato di valore atteso della risposta. Avrà quindi un ruolo notevole a seconda della link function utilizzata.

5.2.1 Identity link: linear probability model

Se utilizziamo la funzione identità come link, otteniamo il *linear probability model*. Per la generica osservazione si ha:

$$\pi(\mathbf{x}_i) = \sum_{j=1}^N \beta_j x_{ij}, \quad i = 1, \dots, N.$$

dove $\mathbf{x}_i = (x_{i1}, \dots, x_{iN})$, il vettore dei predittori del soggetto i .

Un tale modello è semplice da scrivere, ma ha il principale difetto che la probabilità $\pi(\mathbf{x})$ deve cadere tra 0 e 1, mentre per valori sufficientemente grandi di qualche predittore, il valore predetto potrebbe uscire dal range. Se possibile, si può restringere la validità del modello su un opportuno sottoinsieme di predittori \mathbf{x} tali da preservare la consistenza delle predizioni compiute. In tal modo il modello costruito rimane molto semplice e di facile interpretazione.

5.2.2 Canonical link: logistic regression model

Generalmente si preferisce l'utilizzo di una relazione non lineare tra $\pi(\mathbf{x})$ e il vettore dei predittori \mathbf{x} , in cui una variazione fissata dei predittori ha effetti di diversa entità a seconda che il valore di π sia più o meno vicino a 0 o 1, oppure sia vicino a 0,5. Ci aspettiamo che una variazione dei predittori abbia meno peso su una probabilità già molto vicina agli estremi del range consentito piuttosto che su una probabilità vicina a 1/2, ovvero a metà dell'intervallo consentito. Nella pratica si utilizza una relazione che sia monotona crescente (o decrescente).

La relazione con queste caratteristiche che assume il ruolo più importante è la seguente:

$$\pi(\mathbf{x}_i) = \frac{\exp(\sum_{j=1}^N \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^N \beta_j x_{ij})}. \quad (5.5)$$

Questa relazione corrisponde ad utilizzare la funzione di link canonica, ovvero la funzione logit, che corrisponde al parametro naturale della verosimiglianza delle risposte Y :

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \sum_{j=1}^N \beta_j x_{ij}. \quad (5.6)$$

Il modello così specificato è detto di *regressione logistica* o modello logit.

Osserviamo che se π cade tra zero e uno, allora il relativo logit può assumere qualunque valore reale. Il modello logit quindi non è affetto dal problema strutturale del modello linear probability; non è infatti necessario specificare alcun range di validità per i predittori. Per contro i parametri del modello sono di meno immediata interpretazione rispetto al precedente modello.

5.2.3 GLM Binomiale per una risposta contro un predittore binario

Un caso particolare dei precedenti due modelli si può avere quando una variabile risposta Y viene studiata in relazione ad un solo predittore X , anch'esso binario (con valori 0 e 1). In questo caso il significato dell'unico coefficiente β è subito evidente. Per la funzione di link identica si ha $\beta = \pi(1) - \pi(0)$, ovvero la differenza tra le proporzioni per due categorie. Utilizzando la funzione logit come funzione di link si ha:

$$\begin{aligned}\beta &= \log \frac{\pi(1)}{1 - \pi(1)} - \log \frac{\pi(0)}{1 - \pi(0)} \\ &= \log \frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))}\end{aligned}$$

Ovvero, β rappresenta in questo caso il logaritmo del rapporto tra gli *odds ratio* per le due categorie, ovvero qualcosa di molto simile ad un Bayes factor.

5.2.4 Inverse CDF link function: modello probit

Osserviamo che le caratteristiche richieste per la funzione di link (funzione monotona e definita per valori reali in $(0; 1)$) sono verificate perfettamente da una qualsiasi funzione cumulativa inversa. È quindi possibile scrivere un modello del tipo:

$$F^{-1}(\pi(\mathbf{x}_i)) = \sum_{j=1}^N \beta_j x_{ij}$$

Se F rappresenta la funzione di probabilità cumulativa di una Normale di media zero e varianza uno, allora si parla di modello *probit*, per il quale si ha:

$$\Phi^{-1}(\pi(\mathbf{x}_i)) = \sum_{j=1}^N \beta_j x_{ij}.$$

La regressione logistica è un caso particolare, dove la link function è l'inversa della funzione cumulativa di una variabile aleatoria logistica standard. La forma è molto simile alla cumulativa di una Normale standard, con code lievemente più pesanti.

5.3 Media e Varianza per la componente casuale

Scrivendo la verosimiglianza del modello, è possibile ricavare per un GLM l'espressione della media e della varianza delle variabili Y . Per fare questo, proponiamo prima una forma alternativa all'espressione del modello precedentemente scritta nella Formula (5.1):

$$f(y_i; \theta_i, \phi) = \exp\{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}. \quad (5.7)$$

Questa formulazione è chiamata *exponential dispersion family*, e il parametro ϕ è chiamato dispersione. Nel caso particolare di ϕ noto, la precedente formula si può ridurre a quella già espressa nella (5.1). È possibile identificare il precedente termine $Q(\theta)$ con $\theta/a(\phi)$ (di conseguenza θ rappresenta ora il parametro naturale). L'espressione $a(\theta)$ nella (5.1) è ora sostituita da $\exp[-b(\theta)/a(\phi)]$, mentre l'espressione $\exp[c(y, \phi)]$ prende il posto del termine $b(y)$ della (5.1). Generalmente $a(\phi) = \phi/\omega_i$, con ω_i peso noto.

Si può quindi esprimere il logaritmo della verosimiglianza come $l = \sum_i l_i$, dove:

$$l_i = \log(f(y_i; \theta_i, \phi)) = [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi).$$

Di conseguenza:

$$\partial l_i / \partial \theta_i = [y_i - b'(\theta_i)]/a(\phi), \quad \partial^2 l_i / \partial \theta_i^2 = -b''(\theta_i)/a(\phi),$$

dove $b'(\theta_i)$ e $b''(\theta_i)$ rappresentano le prime due derivate di b rispetto a θ_i . Applicando il risultato secondo il quale:

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0, \quad -\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) = \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right)^2$$

che vale sotto opportune condizioni di regolarità per la famiglia esponenziale (condizioni che assumiamo soddisfatte), possiamo ricavare dalla prima formula:

$$\mu_i = \mathbb{E}(Y_i) = b'(\theta_i)$$

e dalla seconda formula:

$$\mathbb{V}(Y_i) = b''(\theta_i)a(\phi).$$

In conclusione la funzione $b(\theta)$ determina i momenti della componente casuale del modello.

Ragionando sulla componente sistematica e sulla link function, usando la link canonica si ha:

$$g(\mu_i) = \theta_i = \sum_j \beta_j x_{ij}.$$

Dal momento che $\mu_i = b'(\theta_i)$, la funzione di link canonica altri non è che l'inversa di $b'(\theta_i)$.

Ad esempio, supponiamo che la variabile $n_i Y_i$ abbia distribuzione Binomiale di parametri n_i e π_i ; di conseguenza la realizzazione y_i rappresenta una proporzione campionaria di successi su n_i prove. Il valore atteso di Y_i resta quindi indipendente da n_i . Definendo $\theta_i = \log[\pi_i/(1 - \pi_i)]$ (e di conseguenza $\pi_i = \exp(\theta_i)/(1 + \exp(\theta_i))$), possiamo scrivere:

$$\begin{aligned} f(y_i; \theta_i, n_i) &= \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - n_i y_i} \\ &= \exp \left[\frac{y_i \theta_i - \log(1 + \exp(\theta_i))}{1/n_i} + \log \binom{n_i}{n_i y_i} \right]. \end{aligned} \quad (5.8)$$

La verosimiglianza è così espressa nella forma della famiglia esponenziale con dispersione indicata nella Formula (5.8), con $b(\theta_i) = \log(1 + \exp(\theta_i))$, $a(\phi) = 1/n_i$ e $c(y_i, \phi) = \log \binom{n_i}{n_i y_i}$. Il parametro naturale è quindi il logit di π_i . Dalle formule ricavate precedentemente abbiamo:

$$\mathbb{E}(Y_i) = b'(\theta_i) = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \pi_i$$

e

$$\mathbb{V}(Y_i) = b''(\theta_i)a(\phi) = \frac{\exp(\theta_i)}{(1 + \exp(\theta_i))^2 n_i} = \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Quella appena presentata come esempio è una forma alternativa molto utile per analizzare risposte binarie. Invece di usare la singola risposta su

una prova, raggruppiamo tutte le n_i prove corrispondenti ai medesimi livelli dei predittori sommando le relative risposte binarie. Passiamo così da una rappresentazione tramite variabili di Bernoulli ad una che sfrutta variabili binomiali. I parametri di interesse rimangono però gli stessi, ovvero i π_i e la loro dipendenza dai predittori. Utilizzando infatti la link canonica ritroviamo il modello logistico, dove il logit di π_i dipende linearmente dai predittori.

5.4 Inferenza e comparazione tra modelli

5.4.1 Deviance

Per un particolare modello generalizzato, date le osservazioni $\mathbf{y} = (y_1, \dots, y_N)$, indichiamo con $l(\boldsymbol{\mu}; \mathbf{y})$ la log-verosimiglianza espressa come funzione della media $\boldsymbol{\mu}$. Con $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ indichiamo il massimo della log-verosimiglianza per il modello. Considerando tutti i possibili modelli, il massimo valore che la log-verosimiglianza potrebbe raggiungere è dato da $l(\mathbf{y}; \mathbf{y})$. Questo valore è raggiunto per il modello più generale possibile, ovvero il modello che ha un parametro diverso per ogni osservazione. Un modello simile presenta un fit perfetto: $\hat{\boldsymbol{\mu}} = \mathbf{y}$. Un tale modello è detto modello *saturato*. Questo modello è di per sé inutile in quanto non riduce la dimensione del problema (spiega infatti i dati con un numero di parametri pari al numero dei dati stessi), ma viene usato come riferimento per giudicare altri modelli.

Possiamo infatti definire la *deviance* come segue:

$$-2[l(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})]. \quad (5.9)$$

Questa è la statistica basata sul rapporto di verosimiglianza per testare l'ipotesi nulla in cui vale il modello fittato contro l'ipotesi alternativa in cui vale un modello più generale, in questo caso il modello saturato. Se il numero di osservazioni resta fisso² a N , allora (per lo meno nel caso di modelli Poisson o Binomiali) la deviance ha sotto H_0 distribuzione asintotica χ_{N-p}^2 ; p rappresenta il numero di parametri del modello lineare. Osserviamo che N è il numero di parametri presenti nel modello saturato, quindi i gradi di libertà non sono altro che la differenza tra il numero di parametri dei due modelli. Quindi la deviance è in grado di evidenziare il *Lack of Fit* del modello costruito.

²Per esempio, in un modello Binomiale, per osservazioni si intendono i diversi possibili setting dei valori dei predittori. Possono invece variare il numero n_i di prove svolte per ogni setting.

Indichiamo con $\tilde{\theta}_i$ la stima del parametro naturale per il modello saturato, che corrisponde alla stima $\tilde{\mu}_i = y_i$, e con $\hat{\theta}_i$ e $\hat{\mu}_i$ le stime di massima verosimiglianza per il modello non saturo. Considerando la verosimiglianza espressa nella formula (5.7), e ricordando che $l = \sum_i l_i$ e che usualmente $a(\phi) = \phi/\omega_i$, possiamo scrivere:

$$\begin{aligned} & -2[l(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] \\ &= 2 \sum_i [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2 \sum_i [y_i \hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi) \\ &= 2 \sum_i \omega_i [y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi. \end{aligned}$$

La quantità $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ è detta *scaled deviance*.

Ad esempio, consideriamo il modello con componente casuale Binomiale, dove le y_i rappresentano delle proporzioni campionarie su n_i prove. Da quanto già visto, sappiamo che $\hat{\theta}_i = \log(\hat{\pi}_i/(1 - \hat{\pi}_i))$ e che $b(\hat{\theta}_i) = -\log(1 - \hat{\pi}_i)$. Similmente abbiamo $\tilde{\theta}_i = \log(y_i/(1 - y_i))$ e $b(\tilde{\theta}_i) = -\log(1 - y_i)$. Ricordiamo anche che $a(\phi) = 1/n_i$. Si può ricavare l'espressione della deviance:

$$2 \sum_i n_i y_i \log \frac{n_i y_i}{n_i \hat{\pi}_i} + 2 \sum_i (n_i - n_i y_i) \log \frac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i}.$$

Dal momento che $\phi = 1$, la deviance corrisponde anche con la scaled deviance.

Come abbiamo già sottolineato, per il modello Binomiale è possibile esprimere i dati come conteggio di successi e insuccessi per ogni setting dei predittori, oppure come singole prove di Bernoulli. Nel primo caso il modello saturato avrà un parametro per ogni setting dei predittori, mentre nel secondo caso il modello saturato avrà un parametro per ogni diverso soggetto. Queste due formulazioni sono chiamate rispettivamente *grouped data* e *un-grouped data*. È possibile assumere una χ^2 come distribuzione asintotica per la deviance solo nel caso di grouped data. In questo caso infatti, al crescere del numero di soggetti studiati, il numero di setting dei predittori può rimanere fisso, e di conseguenza il numero di parametri del modello saturato.

5.4.2 Comparazione di modelli tramite la deviance

La deviance può essere sfruttata per confrontare due modelli di cui uno non sia necessariamente il modello saturato. Consideriamo quindi due modelli, M_0 con valori fittati $\hat{\boldsymbol{\mu}}_0$ e M_1 con valori fittati $\hat{\boldsymbol{\mu}}_1$, in cui M_0 sia un caso speciale di M_1 , ovvero diremo che M_0 è annidato (*nested*) in M_1 .

Essendo M_1 un modello più generale di M_0 , si ha $l(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) \leq l(\hat{\boldsymbol{\mu}}_1; \mathbf{y})$. Segue quindi che $D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) \leq D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0)$. Ciò significa che un modello più semplice potrà avere deviance (o scaled deviance) maggiore di uno più complesso.

La statistica test basata sul rapporto di verosimiglianza per confrontare i due modelli è:

$$\begin{aligned} & -2[l(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}_1; \mathbf{y})] \\ & = -2[l(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] - \{-2[l(\hat{\boldsymbol{\mu}}_1; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})]\} \\ & = D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1). \end{aligned}$$

Da quanto già visto possiamo scrivere:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i \omega_i [y_i(\hat{\theta}_{1i} - \hat{\theta}_{0i}) - b(\hat{\theta}_{1i}) + b(\hat{\theta}_{0i})].$$

La statistica corrisponde quindi alla semplice differenza tra la deviance dei due modelli considerati. Il suo valore cresce al peggiorare del fit di M_0 rispetto a M_1 , e sotto opportune condizioni di regolarità essa ha approssimativamente distribuzione χ^2 con gradi di libertà pari alla differenza tra il numero di parametri dei due modelli.

Si può sottolineare che per il modello Binomiale il confronto tra modelli non risente della scelta tra le formulazioni grouped data o ungrouped data. Questo perchè nonostante il modello saturato differisca tra le due formulazioni, la relativa parte di log-verosimiglianza si elimina per differenza.

5.4.3 Comparazione di modelli tramite l'AIC

Un utile strumento per mettere a confronto diversi modelli di regressione è l'*Aikike Information Criterion*, indicato con AIC. Questo criterio si focalizza sul trade-off tra la complessità del modello e la sua bontà di adattamento.

Dato un certo set di dati \mathbf{y} e un insieme di parametri stimati $\hat{\boldsymbol{\mu}}$, definiamo l'AIC del modello nel seguente modo:

$$AIC = 2k - 2L(\hat{\boldsymbol{\mu}}; \mathbf{y})$$

dove termine k indica il numero di parametri presenti nel modello stimato.

Commentiamo il comportamento di questo indice. Un modello con fit migliore avrà una log-verosimiglianza più grande, e di conseguenza un valore di AIC minore. Per ottenere modelli con miglior fitting però una strada

possibile è quella di aumentare indiscriminatamente il numero dei parametri; ciò porta ad avere modelli con log-verosimiglianza più alta, ma questi stessi modelli hanno minore utilità in quanto diventano troppo complessi. Per questo il termine $2k$ viene inserito per penalizzare una complessità troppo elevata del modello. Ecco quindi che se si ha a disposizione una famiglia di modelli, un'idea per confrontarli è osservare il valore dei rispettivi AIC scegliendo quello che presenta il valore inferiore.

D'altro canto però, da un punto di vista generale, l'AIC non permette di svolgere un test vero e proprio, ma ci mette a disposizione solamente un indice informativo della bontà del modello. Nel caso particolare di modelli con lo stesso numero di parametri, il confronto tra gli AIC può essere interpretato come un test sul rapporto di verosimiglianza, se i modelli confrontati sono modelli annidati l'uno nell'altro.

Un altro punto a favore dell'indice AIC è che può essere usato anche per confrontare modelli non necessariamente annidati.

5.4.4 Residui per i GLM

Anche per i modelli lineari generalizzati può essere utile l'analisi dei residui del modello per osservare eventuali outlier o problemi di fitting. In generale però queste procedure sono meno precise e sviluppate rispetto alle tecniche diagnostiche per i modelli lineari classici.

Tipologie di residui

Sono possibili diverse definizioni di residui. Cominciamo con l'osservare che si può scrivere $D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum d_i$, dove $d_i = 2\omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$. Si definiscono quindi i *deviance residuals* come:

$$\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i).$$

La somma di questi residui elevati al quadrato danno di conseguenza la deviance del modello.

Una seconda possibilità sono i *Pearson residuals*:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(Y_i)}}.$$

La somma dei residui di Pearson elevati al quadrato equivale a X^2 , la statistica del test di Pearson che si distribuisce asintoticamente come una χ^2_{N-p} .

La terza tipologia di residui sono gli *standardized residuals*, che corrispondono ai residui ordinari divisi per il loro standard. Per un GLM abbiamo:

$$\text{Cov}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \text{Cov}(\mathbf{Y})[\mathbf{I} - \mathbf{H}].$$

Con \mathbf{I} indichiamo la matrice identità e con \mathbf{H} la matrice *Hat*, la matrice che applicata al vettore dei dati \mathbf{y} fornisce il vettore delle previsioni $\hat{\boldsymbol{\mu}}$. In particolare si può dimostrare che:

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}^{1/2},$$

dove \mathbf{W} è una matrice diagonale, le cui componenti non nulle sono definite come segue:

$$w_{ii} = \frac{\partial \mu_i}{\partial \eta_i} \times \frac{1}{\mathbb{V}(Y_i)}.$$

In particolare per i modelli di regressione gaussiani, si avrebbe $w_i = 1/\sigma^2$, dove σ^2 è la varianza di Y_i . Riotteniamo così la ben nota relazione per i modelli di regressione classica:

$$\mathbf{H} = \frac{1}{\sigma} \mathbf{X} \left(\frac{1}{\sigma^2} \mathbf{X}^t \mathbf{X} \right)^{-1} \mathbf{X}^t \frac{1}{\sigma} = \mathbf{X} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t.$$

Il residuo standardizzato per l'osservazione i -esima diventa quindi:

$$r_i = \frac{y_i - \hat{\mu}_i}{\hat{\mathbb{V}}(Y_i) \sqrt{1 - \hat{h}_i}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}.$$

La quantità \hat{h}_i rappresenta la stima della componente diagonale in posizione i della matrice \mathbf{H} . Questa quantità è chiamata *leverage*, come già per i modelli di regressione consueti. Una osservazione con alto leverage è una osservazione potenzialmente influente sulla stima dei coefficienti del modello. Così come per la regressione lineare, i leverage sono compresi tra zero e uno, e sommano al numero di parametri del modello.

Osserviamo che, a differenza di quanto avveniva per i modelli lineari classici, la matrice \mathbf{H} dipende attraverso \mathbf{W} dalla stima dei coefficienti. Di conseguenza non è possibile valutare la matrice \mathbf{H} prima della raccolta dei dati, come invece si potrebbe fare in fase di progettazione dell'esperimento usando una regressione lineare classica.

Brier's score

Per i residui provenienti dalla regressione lineare classica è possibile trarre diverse informazioni dai loro grafici. Un esempio su tutti, dal grafico dei residui contro i valori predetti si può verificare visivamente l'assenza di particolari pattern nella nuvola ottenuta. Ciò è indizio che i residui sono *approssimativamente* indipendenti dalle previsioni; nella regressione lineare classica ci si aspetta un comportamento del genere in quanto i residui sono delle quantità piuttosto vicine agli errori.

Per i residui provenienti da un modello di regressione generalizzato, tali interpretazioni sono piuttosto complesse. Questi residui non hanno più una tale utilità diretta.

Nonostante questo, è possibile utilizzare un indice di bontà legato ai residui, chiamato **Brier's score**. Esso è definito come segue:

$$BS = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

Questa quantità è molto simile alla somma dei quadrati degli errori presente in una tabella ANOVA: y_i rappresenta la quantità osservata (ad esempio una variabile binaria) mentre con \hat{y}_i indichiamo la relativa previsione. Più l'indice cresce, più i residui³ hanno valori maggiori. È preferibile quindi un modello che presenti un BS minore.

Questo indice, con leggere modifiche, è particolarmente utilizzato per giudicare la bontà di previsioni di eventi: in particolar modo è molto presente nella letteratura riguardante la previsione di eventi di natura meteorologica. Fu proposto da Glenn W. Brier nel 1950.

5.5 La Regressione Logistica

Richiamiamo quanto già detto poco indietro: per una variabile Y con valori 0 o 1 e un predittore⁴ X , indichiamo con $\pi(x) = \mathbb{P}(Y = 1|X = x)$. Per questa quantità viene utilizzato il modello logistico:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

o equivalentemente

$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x.$$

³Data la definizione di BS , ci riferiamo ai residui di Pearson moltiplicati per $\hat{V}(Y_i)$.

⁴Successivamente ci occuperemo del caso con più predittori. Ci limitiamo per ora ad un predittore per ragioni di semplicità.

Ricordiamo anche che la funzione logistica corrisponde in questo caso alla funzione di link canonica.

5.5.1 Interpretazione dei coefficienti e forma della curva

Per prima cosa osserviamo che il segno di β determina il senso di crescita di $\pi(x)$: se $\beta > 0$, $\pi(x)$ cresce al crescere di x , mentre decresce con x se vale che $\beta < 0$. Se $\beta = 0$, allora Y è indipendente da X . Un aumento di una unità in x determina che il rapporto $\frac{\pi_i}{1-\pi_i}$ aumenti di un fattore moltiplicativo e^β . È quindi legato all'inclinazione della curva logistica e alla sua forma.

Usando una approssimazione lineare della curva logistica, si può verificare che vicino al valore di x per cui $\pi(x) = 1/2$ la derivata della curva logistica vale $\beta/4$; perciò una variazione di x di una quantità $1/\beta$ corrisponde ad una variazione di $\pi(x)$ di circa $1/4$. La quantità $1/\beta$ approssima quindi la distanza tra i valori di x in cui si $\pi(x) = 0.25$ (o $\pi(x) = 0.75$) e il valore di x in cui si ha $\pi(x) = 0.5$. Quest'ultimo corrisponde a $x = -\alpha/\beta$, chiamato anche *median effective level*⁵, e denotato con EL_{50} .

Per quanto riguarda il parametro α , questo non è molte volte di particolare interesse. Osserviamo che centrando i valori del predittore intorno alla sua media (ad esempio sostituendo x con $(x - \bar{x})$, dove \bar{x} è la media di X), α assume il valore di logit nella media; di conseguenza $\pi(\bar{x}) = e^\alpha/(1 + e^\alpha)$.

In Figura 5.1 riportiamo il grafico della funzione logistica per alcuni valori di α e β . È riportata anche la retta che approssima la funzione nel punto $x = -\alpha/\beta$ e $\pi(x) = 1/2$. Si può osservare che (tenendo conto della scala dei grafici), al diminuire di β , la pendenza nel punto EL_{50} diminuisce, mentre all'aumentare di α , questo punto si sposta verso valori negativi. Osserviamo che, data la simmetria della funzione logistica, il punto denominato EL_{50} è anche un punto di flesso.

5.5.2 Inferenza

Così come per la regressione lineare classica, è possibile valutare il modello costruito attraverso diversi strumenti. Per prima cosa mostriamo alcuni test di significatività per le stime dei coefficienti. Successivamente vediamo come investigare la bontà di adattamento del modello.

⁵O *lethal dose* in studi di tossicologia; indica generalmente la dose di sostanza cui corrisponde una probabilità di $1/2$ di sopravvivere.

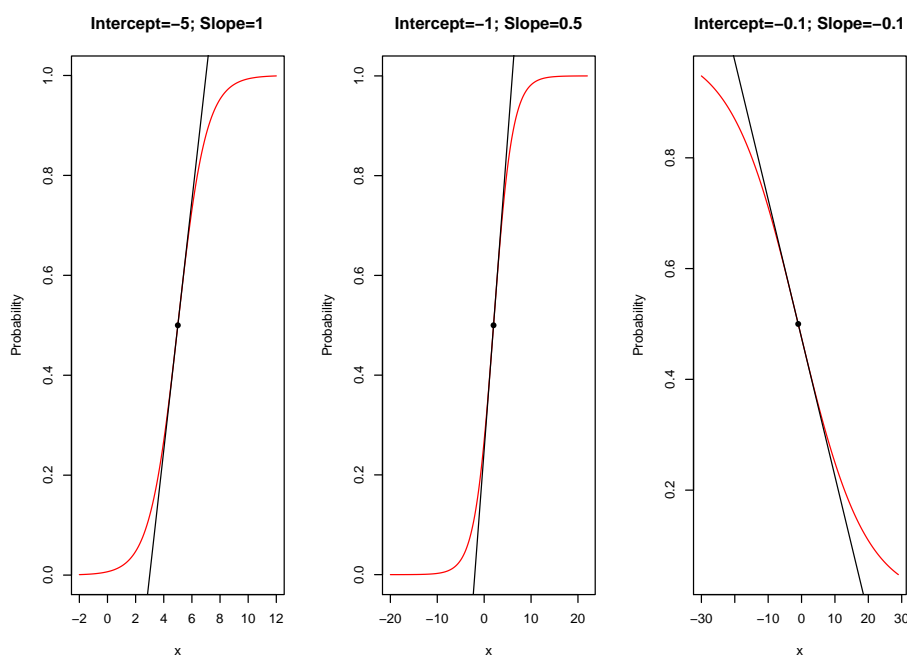


Figura 5.1: Grafici per la funzione logistica per alcuni valori di α e β .

Test di significatività

Una volta stimati i coefficienti (generalmente attraverso le equazioni di massima verosimiglianza), è possibile fare opportune operazioni inferenziali per giudicare il modello costruito.

I test di significatività si focalizzano su $H_0 : \beta = 0$ contro $H_1 : \beta \neq 0$. Diversi sono i possibili approcci. Un risultato asintotico dovuto a Wald ci garantisce che la stima di massima verosimiglianza $\hat{\beta}$ è approssimativamente normale, oltre ad essere uno stimatore non distorto di β . Di conseguenza z^2 , dove $z = \hat{\beta}/SE$, è approssimativamente una χ_1^2 . Il test basato sul rapporto di verosimiglianza sfrutta invece la differenza tra il logaritmo della verosimiglianza calcolato in $\hat{\beta}$ e in $\beta = 0$. Anche in questo caso la distribuzione di riferimento è una χ_1^2 . Fra questi test, generalmente il test sul rapporto di verosimiglianza è preferito al test di Wald, prima di tutto in quanto sfrutta maggior informazioni e soprattutto in quanto è più potente, in particolare per valori di β grandi. Infatti, all'aumentare del valore di β , aumenta anche lo standard error utilizzato nella statistica di Wald, incrementando la probabilità di errore del secondo tipo di questo test.

Un altro strumento molto informativo è costituito dagli intervalli di confidenza, che possono essere costruiti invertendo un test per $H_0 : \beta = \beta_0$.

Questo intervallo è l'insieme dei valori di β_0 per cui la statistica chi-quadro non supera $\chi_1^2(\alpha) = z_{\alpha/2}^2$. Con l'approccio di Wald, l'intervallo diventa $\beta \in [\hat{\beta} \pm z_{\alpha/2} \times (SE)]$.

Si possono ricavare stime anche per il logit di $\pi(x)$. Per un valore fissato di x , ad esempio x_0 , il logit di $\pi(x_0)$ ha una stima asintotica dello standard error data dalla radice quadrata di:

$$\mathbb{V}(\hat{\alpha} + \hat{\beta}x_0) = \mathbb{V}(\hat{\alpha}) + x_0^2\mathbb{V}(\hat{\beta}) + 2x_0\text{Cov}(\hat{\alpha}, \hat{\beta}).$$

Quindi l'intervallo $(\hat{\alpha} + \hat{\beta}x_0) \pm 1.96(SE)$ è un insieme di confidenza al 95% per il logit di $\pi(x_0)$. È possibile anche ricavare un intervallo di confidenza per $\pi(x_0)$ invertendo la relazione tra $\pi(x_0)$ e $\log \pi(x_0)/(1 - \pi(x_0))$.

Goodness of Fit

Nella pratica non vi è certezza che un modello spieghi bene i dati. Per verificare il possibile Lack of Fit è possibile confrontare il modello costruito con modelli più complessi. Se modelli più complicati non migliorano sostanzialmente il fit dei dati, vi è abbastanza sicurezza del fatto che il modello più semplice sia ragionevole. Ad esempio si può compiere questo confronto rispetto al modello saturato, come già indagato in precedenza, utilizzando quindi la deviance. In questo caso è importante la scelta tra forma raggruppata e non raggruppata del modello. Nella forma raggruppata, il numero di setting dei valori dei predittori resta fissato; all'aumentare delle dimensioni campionarie n_i , la deviance ha una distribuzione approssimativamente χ^2 , con gradi di libertà (detti gradi di libertà residui) pari alla differenza tra il numero di parametri del modello saturato e il numero di parametri del modello più semplice. Facendo un confronto tra due modelli annidati (quali ad esempio il modello corrente e il modello nullo, cioè con $\beta = 0$), la distinzione tra modello raggruppato o non raggruppato perde importanza.

In alternativa, si possono confrontare le probabilità predette dal modello con le frequenze osservate tramite un test X^2 di buon adattamento di Pearson, o con la sua variante G^2 . Per queste statistiche si ha, al crescere delle probabilità predette dal modello e per un numero di gruppi fissato, una distribuzione approssimativamente χ^2 . Ancora una volta i gradi di libertà sono pari al numero di gruppi meno il numero di parametri del modello. Questi test potrebbero però non rispettare le opportune ipotesi per ottenere la distribuzione asintotica desiderata: ad esempio le probabilità fittate potrebbero essere molto piccole. Ciò può avvenire quando si hanno molti diversi livelli per i valori dei predittori, ad esempio quando il predittore è continuo. Quello che è possibile fare è raggruppare i dati in un numero minore di categorie,

suddividendo sostanzialmente lo spazio dei possibili valori dei predittori. In tal maniera si elimina il problema delle stime troppo piccole. Si deve tenere conto però che in ogni categoria, la probabilità non sarà omogenea; oltretutto quando si ha a che fare con molti predittori una tale procedura può rivelarsi molto complessa. Con pochi predittori (ad esempio uno solo) la suddivisione in classi è più agevole.

5.5.3 Regressione logistica con un predittore categoriale

Anche nella regressione logistica è possibile avere a che fare con predittori che assumono una quantità discreta di possibili valori. Essi sono detti *fattori*, così come nella regressione ordinaria, o meglio nell'analisi fattoriale e ANOVA. Nel caso di un unico fattore X con I possibili livelli, i dati sono esprimibili come una tabella $I \times 2$, dove la prima colonna contiene le y_i , che rappresenta il numero di successi su n_i prove. Il modello logistico diventa quindi:

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_i$$

con $i = 1 \cdots I$. Come nell'ANOVA tradizionale, uno dei parametri per X è ridondante, e si può riscalarlo il vettore β in modo che uno dei parametri valga zero, per esempio β_1 . In tal modo α rappresenta il logit quando $X = 1$. Questo modello ha tanti parametri quanti il numero di osservazioni binomiali, e di conseguenza è saturato. Quando il fattore non ha effetto sulla probabilità, si ha $\beta_1 = \beta_2 = \cdots = \beta_I$. In tal caso $\pi_1 = \pi_2 = \cdots = \pi_I$, e X e Y sono indipendenti.

Il modello precedente può essere espresso utilizzando delle variabili dummy (o variabili indicatore) x_i , con i tra 1 e $I - 1$; avremo che $x_i = 1$ per le osservazioni al livello i , altrimenti $x_i = 0$. Il modello diventa quindi:

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{I-1} x_{I-1}.$$

La forma appena presentata corrisponde al vincolo $\beta_I = 0$, e non introduce un variabile dummy per la categoria I . Osserviamo che un'altra possibilità sarebbe utilizzare il vincolo $\sum_i \beta_i = 0$. Per un fattore a due livelli questo vincolo determina, per esempio, che $\beta_1 = -\beta_2$; le variabili dummy diventano quindi due come il numero di livelli del fattore e si avrebbe $x_1 = 1$ e $x_2 = -1$. La variazione del vincolo imposto determina diversi valori dei coefficienti, ma non crea variazioni nelle stime $\hat{\pi}_i$, così come nelle differenze tra le stime delle diverse categorie.

5.6 Regressione logistica multipla

Nel caso siano presenti più predittori per il logit di π_i , il modello di regressione logistica diventa:

$$\text{logit}[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Ora β_i rappresenta l'impatto del predittore x_i sul logit di π a parità di condizioni degli altri predittori: tenendo fissati gli altri predittori, un aumento di una unità nel valore di x_i determina la variazione secondo un fattore moltiplicativo e^{β_i} sul valore del rapporto $\pi/(1 - \pi)$.

Come per il caso di un solo predittore, alcuni dei predittori possono essere dei fattori, introducendo delle variabili dummy per descriverne le categorie.

Come nei modelli di regressione classica o nei modelli di analisi fattoriale, è possibile includere anche la possibile interazione tra i predittori (fattori) in gioco, rendendo il modello più raffinato per cogliere l'influenza che l'appartenenza ad una particolare categoria di un predittore può indurre sulla relazione tra la risposta e gli altri predittori. In particolare l'assenza di interazione tra predittori ci dice che l'influenza di un predittore sulla risposta non dipende dal particolare valore di un altro predittore.

Per verificare il buon fitting del modello, ancora una volta è possibile sfruttare la statistica basata sul rapporto di verosimiglianza per testare se alcuni parametri del modello sono pari a zero. Nell'ipotesi nulla viene di conseguenza posto un modello ridotto M_0 : per esempio se M_0 è il modello nullo, il test ci da una visione complessiva della significatività del modello fittato. Come prima, si può mostrare che la statistica basata sul rapporto di verosimiglianza per confrontare due modelli annidati equivale alla differenza tra le deviance dei modelli stessi. Questa quantità, sotto opportune ipotesi, si distribuisce asintoticamente come una χ^2 con gradi di libertà pari alla differenza tra il numero dei parametri nei due modelli.

Sottolineiamo nuovamente che in alcuni casi, nonostante la deviance del modello non abbia una distribuzione χ^2 , la statistica per il confronto di due modelli, pari alla differenza tra le rispettive deviance, ha comunque una distribuzione approssimativamente χ^2 ; ciò in particolare quando i modelli confrontati differiscono per pochi parametri.

5.7 Analisi dei dati del progetto *Scompenso*

Vogliamo applicare quanto visto in questo capitolo riguardo alla regressione logistica e alla regressione logistica multipla ai dati presentati nel Capitolo

4. In particolare vogliamo spiegare la probabilità di morte all'interno dell'ospedale in base ad opportune covariate. Utilizzeremo quindi un modello di regressione logistica multipla.

Cominciamo col ricordare che, dopo opportune operazioni di pulizia del data set, ci siamo ridotti ad un numero di pazienti pari a 26650. Di questi pazienti, 13414 sono femmine mentre 13236 sono maschi. Per ogni paziente consideriamo le seguenti covariate di interesse:

- sesso;
- età (relativa all'ultimo ricovero);
- il numero delle comorbidità accumulate all'ultimo ricovero, che chiamiamo *indice di salute* (`hidx`);
- le variabili SHOCK, CABG, PTCA, STENT, ICD che indicano quante volte il paziente, nella sua storia clinica, ha subito (rispettivamente) eventi di shock cardiogeno, bypass coronarici, interventi di angioplastica, impianti di stent cardiaci e trattamenti con *implantable cardiac defibrillator*;
- il numero di ricoveri subiti dal paziente nel periodo dal 2006 al 2010 (`npr`);
- l'identificativo dell'ospedale dell'ultima ospedalizzazione subita.

Osserviamo che dopo la riduzione del data set alla forma sopra indicata, il numero degli ospedali cui viene attribuito almeno un paziente ammonta a 412 unità rispetto alle 497 riportate nei Funnel Plot.

Prima di proseguire, osserviamo che le variabili età e sesso sono molto probabilmente dipendenti; questo perchè le femmine hanno generalmente un'età di insorgenza di problemi cardio-vascolari più avanzata rispetto ai maschi, o comunque tendono ad essere ricoverate ad età più elevate. In ogni caso non sembra essere grave questo effetto e di conseguenza lo trascureremo.

Utilizzeremo per semplicità modelli additivi, senza interazioni, essenzialmente per motivi di semplicità di interpretazione del modello. Per quanto riguarda le variabili SHOCK, PTCA, ICD, CABG e STENT cominciamo con l'inserire tra i regressori indicatori del tipo $I(\text{SHOCK} \neq 0)$, che indicano se la variabile considerata è uguale o diversa da zero. Lo facciamo essenzialmente per semplicità; successivamente proveremo con l'inserimento diretto dei fattori.

Cominciamo con un modello additivo in cui entrino tutte le variabili considerate:

```
fit_ok1<-glm(dead_intraH ~ age + sex + hidx + npr + I(SHOCK!=0)
             + I(PTCA!=0) + I(ICD!=0) + I(CABG!=0) + I(STENT!=0),
             data=msm, family='binomial')
```

Per questo modello si può osservare la tabella dei coefficienti con i relativi test di Wald:

```
> summary(fit_ok1)
```

Call:

```
glm(formula = dead_intraH ~ age + sex + hidx + npr + I(SHOCK !=
  0) + I(PTCA != 0) + I(ICD != 0) + I(CABG != 0) + I(STENT !=
  0), family = "binomial", data = msm)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.4975	-0.6954	-0.5743	-0.3236	3.1722

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.916560	0.142172	-34.582	< 2e-16 ***
age	0.041415	0.001752	23.634	< 2e-16 ***
sex1	-0.236175	0.034221	-6.901	5.15e-12 ***
hidx	0.071800	0.013092	5.484	4.16e-08 ***
npr	0.028109	0.006812	4.127	3.68e-05 ***
I(SHOCK != 0)TRUE	3.542732	0.095896	36.944	< 2e-16 ***
I(PTCA != 0)TRUE	-0.279569	0.202202	-1.383	0.16678
I(ICD != 0)TRUE	-0.941116	0.123448	-7.624	2.47e-14 ***
I(CABG != 0)TRUE	0.691518	0.261307	2.646	0.00814 **
I(STENT != 0)TRUE	-0.355294	0.214130	-1.659	0.09707 .

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27538 on 26649 degrees of freedom
 Residual deviance: 24110 on 26640 degrees of freedom
 AIC: 24130

Tutte le covariate sono indicate come significative, tranne STENT e PTCA. Da questa tabella, come abbiamo già osservato, è complesso capire quali

fattori siano effettivamente significativi, innanzi tutto perchè questi sono test marginali, e anche perchè il test di Wald è affetto da una certa mancanza di potenza. Osserviamo quindi la tabella di analisi della devianza:

```
> anova(fit_ok1, test='Chi')
Analysis of Deviance Table

Model: binomial, link: logit

Response: dead_intraH

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			26649	27538	
age	1	803.62	26648	26734	< 2.2e-16 ***
sex	1	60.05	26647	26674	9.234e-15 ***
hidx	1	82.58	26646	26591	< 2.2e-16 ***
npr	1	12.92	26645	26579	0.0003252 ***
I(SHOCK != 0)	1	2323.16	26644	24255	< 2.2e-16 ***
I(PTCA != 0)	1	63.76	26643	24192	1.403e-15 ***
I(ICD != 0)	1	72.18	26642	24119	< 2.2e-16 ***
I(CABG != 0)	1	6.39	26641	24113	0.0114679 *
I(STENT != 0)	1	2.67	26640	24110	0.1019909

```
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Da questa tabella possiamo effettivamente notare che il termine `I(STENT != 0)` è poco significativo; non c'è evidenza a favore della sua aggiunta al modello costruito con gli altri regressori. Possiamo quindi vedere cosa accade eliminando questo termine dal modello.

Costruiamo quindi un secondo modello:

```
fit_ok2<-glm(dead_intraH ~ age + sex + hidx + npr + I(SHOCK!=0)
              + I(PTCA!=0) + I(ICD!=0) + I(CABG!=0),
              data=msm, family='binomial')
```

da cui abbiamo levato l'effetto di `I(STENT != 0)`. Osserviamo la tabella di significatività dei coefficienti

```
> summary(fit_ok2)
```

```
Call:
```

```
glm(formula = dead_intraH ~ age + sex + hidx + npr + I(SHOCK !=
  0) + I(PTCA != 0) + I(ICD != 0) + I(CABG != 0), family = "binomial",
  data = msm)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-2.4980	-0.6952	-0.5741	-0.3248	3.1737

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.921115	0.142160	-34.617	< 2e-16 ***
age	0.041466	0.001752	23.664	< 2e-16 ***
sex1	-0.236554	0.034220	-6.913	4.75e-12 ***
hidx	0.071859	0.013093	5.488	4.06e-08 ***
npr	0.028139	0.006812	4.131	3.61e-05 ***
I(SHOCK != 0)TRUE	3.543325	0.095864	36.962	< 2e-16 ***
I(PTCA != 0)TRUE	-0.590016	0.082377	-7.162	7.93e-13 ***
I(ICD != 0)TRUE	-0.943456	0.123508	-7.639	2.19e-14 ***
I(CABG != 0)TRUE	0.692811	0.261415	2.650	0.00804 **

```
---
```

```
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 27538 on 26649 degrees of freedom
Residual deviance: 24113 on 26641 degrees of freedom
AIC: 24131
```

Tutti sono indicati come significativi all'1%. Possiamo osservare che il valore dell'AIC è pressochè identico tra i due modelli (24130 contro 24131). Osserviamo la scomposizione della devianza:

```
> anova(fit_ok2, test='Chi')
```

```
Analysis of Deviance Table
```

```
Model: binomial, link: logit
```

```
Response: dead_intraH
```

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			26649	27538	
age	1	803.62	26648	26734	< 2.2e-16 ***
sex	1	60.05	26647	26674	9.234e-15 ***
hidx	1	82.58	26646	26591	< 2.2e-16 ***
npr	1	12.92	26645	26579	0.0003252 ***
I(SHOCK != 0)	1	2323.16	26644	24255	< 2.2e-16 ***
I(PTCA != 0)	1	63.76	26643	24192	1.403e-15 ***
I(ICD != 0)	1	72.18	26642	24119	< 2.2e-16 ***
I(CABG != 0)	1	6.39	26641	24113	0.0114679 *

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Questa tabella è chiaramente uguale alla precedente; semplicemente manca l'ultima riga relativa alla variabile **STENT**. Possiamo vedere che tutti i termini sono significativi almeno al 5%. Il modello sembrerebbe quindi abbastanza buono. Anche levando l'effetto di **CABG**, il modello non migliora, anzi l'AIC cresce (anche se non in maniera drastica):

```
> fit_ok3<-glm(dead_intraH ~ age + sex + hidx + npr +
+ I(SHOCK!=0) + I(PTCA!=0) + I(ICD!=0), data=msm, family='binomial')
> AIC(fit_ok3)
[1] 24135.37
```

Un'altra possibile modifica al modello sarebbe quello di discretizzare l'indice **hidx** in categorie di differente gravità di salute. Per esempio possiamo creare un indice di salute modificato nel seguente modo:

```
hidx_m<-rep(0, dim(msm)[1])

for(i in 1:dim(msm)[1]){
  if (msm$hidx[i]>=3 && msm$hidx<=5) {hidx_m[i]<-1}
  if (msm$hidx[i]>=6 && msm$hidx<=8) {hidx_m[i]<-2}
  if (msm$hidx[i]>=9 && msm$hidx<=11) {hidx_m[i]<-3}
}
```

Di conseguenza l'indice modificato vale zero se **hidx** è minore di 3 (categoria meno grave), vale 1 se **hidx** è compreso tra 3 e 5, vale 2 se **hidx** cade tra 6 e 8,

e vale 3 negli altri casi (categoria di gravità massima). I valori di soglia delle quattro categorie sono scelti in questo modo per motivi di bilanciamento. Il termine `hidx` varia infatti tra zero e undici; assume quindi dodici possibili valori. Ogni categoria raccoglie tre di questi.

Inserendolo come `factor` nella funzione `glm` otterremmo:

```
fit_ok4<-glm(dead_intraH ~ age + sex + as.factor(hidx_m) + npr
             + I(SHOCK!=0) + I(PTCA!=0) + I(ICD!=0) + I(CABG!=0),
             data=msm, family='binomial')
```

Osserviamo la tabella dei coefficienti:

```
> summary(fit_ok4)
```

Call:

```
glm(formula = dead_intraH ~ age + sex + as.factor(hidx_m) + npr +
     I(SHOCK != 0) + I(PTCA != 0) + I(ICD != 0) + I(CABG != 0),
     family = "binomial", data = msm)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4587	-0.6945	-0.5759	-0.3268	3.1631

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.822410	0.141418	-34.100	< 2e-16 ***
age	0.041765	0.001749	23.884	< 2e-16 ***
sex1	-0.241614	0.034189	-7.067	1.58e-12 ***
as.factor(hidx_m)1	0.110455	0.037022	2.984	0.00285 **
as.factor(hidx_m)2	0.299999	0.073741	4.068	4.74e-05 ***
as.factor(hidx_m)3	0.736078	0.400746	1.837	0.06624 .
npr	0.034094	0.006548	5.207	1.92e-07 ***
I(SHOCK != 0)TRUE	3.538913	0.095811	36.936	< 2e-16 ***
I(PTCA != 0)TRUE	-0.598179	0.082348	-7.264	3.76e-13 ***
I(ICD != 0)TRUE	-0.954460	0.123516	-7.727	1.10e-14 ***
I(CABG != 0)TRUE	0.693590	0.261294	2.654	0.00794 **

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 27538 on 26649 degrees of freedom
Residual deviance: 24123 on 26639 degrees of freedom
AIC: 24145

```

Notiamo subito il valore del AIC, pari a 24145. Il modello quindi non migliora rispetto a quelli precedenti. È quindi preferibile utilizzare l'indice `hidx` piuttosto che la sua versione modificata. Possiamo comunque osservare la tabella di analisi della devianza, da cui emergerebbe che i termini inseriti sono tutti significativi al 5%:

```

> anova(fit_ok4, test='Chi')
Analysis of Deviance Table

Model: binomial, link: logit

Response: dead_intraH

Terms added sequentially (first to last)


```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			26649	27538	
age	1	803.62	26648	26734	< 2.2e-16 ***
sex	1	60.05	26647	26674	9.234e-15 ***
as.factor(hidx_m)	3	64.69	26644	26609	5.857e-14 ***
npr	1	21.07	26643	26588	4.423e-06 ***
I(SHOCK != 0)	1	2318.73	26642	24270	< 2.2e-16 ***
I(PTCA != 0)	1	65.91	26641	24204	4.711e-16 ***
I(ICD != 0)	1	74.09	26640	24130	< 2.2e-16 ***
I(CABG != 0)	1	6.41	26639	24123	0.01138 *

```

---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

```

Consideriamo quindi per ora il secondo modello che abbiamo riportato e ragioniamo sui valori delle stime dei coefficienti. Il coefficiente relativo all'età ha segno positivo, per cui un aumento dell'età determina un aumento della probabilità di morte intra ospedaliera. Il fattore sesso al livello 1 (femmine) determina invece un effetto negativo, evidenziando una minore predisposizione alla morte intra ospedaliera per le pazienti donne. I termini `npr` e `hidx` hanno invece un effetto positivo, al loro aumentare, sulla probabilità di morte dei pazienti, così come evidenziano una probabilità di morte maggiore

quei pazienti che abbiano presentato almeno un evento di shock cardiogeno o almeno un intervento di bypass coronarico. Per concludere, hanno una probabilità minore di morte intra ospedaliera quei pazienti che sono stati sottoposti almeno una volta ad un trattamento con ICD o ad un intervento di angioplastica.

Proviamo a modificare il secondo modello fittato inserendo direttamente i fattori SHOCK, PTCA, ICD e CABG. Otteniamo quanto segue:

```
> fit_ok5<-glm(dead_intraH ~ age + sex + hidx +
               npr + SHOCK + PTCA + ICD + CABG,
               data=msm, family='binomial')
> summary(fit_ok5)
Call:
glm(formula = dead_intraH ~ age + sex + hidx + npr + SHOCK +
     PTCA + ICD + CABG, family = "binomial", data = msm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4664  -0.6957  -0.5759  -0.3210   3.1706

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.913674    0.141886 -34.631 < 2e-16 ***
age          0.041383    0.001750  23.649 < 2e-16 ***
sex1        -0.235164    0.034193  -6.878 6.09e-12 ***
hidx        0.070454    0.013094   5.380 7.43e-08 ***
npr         0.029295    0.006823   4.293 1.76e-05 ***
SHOCK       3.482332    0.094567  36.824 < 2e-16 ***
PTCA       -0.489724    0.066685  -7.344 2.07e-13 ***
ICD        -0.918419    0.119376  -7.694 1.43e-14 ***
CABG       0.701042    0.261133   2.685 0.00726 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27538  on 26649  degrees of freedom
Residual deviance: 24133  on 26641  degrees of freedom
AIC: 24151

Number of Fisher Scoring iterations: 5
```

Osserviamo anche la tabella di decomposizione della devianza:

```
> anova(fit_ok5, test='Chi')
Analysis of Deviance Table

Model: binomial, link: logit

Response: dead_intraH

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			26649	27538	
age	1	803.62	26648	26734	< 2.2e-16 ***
sex	1	60.05	26647	26674	9.234e-15 ***
hidx	1	82.58	26646	26591	< 2.2e-16 ***
npr	1	12.92	26645	26579	0.0003252 ***
SHOCK	1	2292.06	26644	24286	< 2.2e-16 ***
PTCA	1	70.15	26643	24216	< 2.2e-16 ***
ICD	1	76.32	26642	24140	< 2.2e-16 ***
CABG	1	6.56	26641	24133	0.0104309 *

```
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Le stime dei coefficienti sono molto simili a quelle precedenti, preservando lo stesso significato appena visto sopra, ma notiamo che l'AIC è più elevato (24131 contro 24151). Questo modello sembrerebbe meno preferibile rispetto al precedente. Possiamo anche confrontare i Brier's Score dei due modelli:

```
> mean((msm$dead_intraH - fitted.values(fit_ok2))^2)
[1] 0.1425657
> mean((msm$dead_intraH - fitted.values(fit_ok5))^2)
[1] 0.142559
```

I valori sono molto simili, anche se il secondo è lievemente più basso. Scegliamo di conseguenza il modello più semplice, quello in cui i termini SHOK, PTCA, ICD e CABG sono inseriti come indicatori binari.

Nel capitolo successivo proveremo a modificare questo modello, aggiungendo un effetto casuale di raggruppamento dovuto all'ospedale, nel tentativo di individuare quegli ospedali caratterizzati dagli effetti più significativi. Per fare questo dovremo presentare gli elementi basilari dei GLMM.

Capitolo 6

Generalized Linear Mixed Models

NEL Capitolo 5 abbiamo introdotto i modelli lineari generalizzati come una estensione alle risposte categoriche delle tecniche di regressione lineare. Abbiamo poi presentato il modello di regressione logistica per dati binari. Quello che vogliamo fare ora è presentare una ulteriore generalizzazione, partendo da una constatazione: che molto spesso i dati si presentano in forma clusterizzata, ovvero in gruppi all'interno dei quali le osservazioni provengono da condizioni omogenee e tendono di conseguenza ad essere maggiormente simili rispetto ad osservazioni provenienti da gruppi differenti. Questo porta alla creazione di una forma di dipendenza tra le osservazioni, infatti risposte provenienti dal medesimo cluster tendono ad essere tra loro positivamente correlate.

I modelli che tengono conto di questi effetti di cluster sono detti *random effect*¹, i quali generalmente introducono dei termini legati all'appartenenza dell'osservazione ad un particolare gruppo: osservazioni nello stesso gruppo hanno un termine di cluster uguale, che differisce da quello cui sono soggette le osservazioni di un altro gruppo.

Nel caso di modelli per risposte non gaussiane, si parla di *Generalized Linear Mixed Models* (GLMM), di cui il modello più importante per dati binari è il *logistic-normal model*.

Come per il precedente, il materiale di questo capitolo è essenzialmente contenuto in [1], al capitolo 12.

¹Abbiamo già avuto la possibilità di usare un semplice modello a fattore casuale nel Capitolo 3 in merito alla costruzione dei Funnel Plot.

6.1 Elementi di un GLMM

In un modello lineare generalizzato ad effetti misti sono presenti sia i parametri che descrivono l'effetto dei fattori e dei predittori sulla risposta, detti *fixed effects*, sia dei parametri di cluster, detti *random effects*, i quali si applicano a tutti i dati appartenenti ad un certo gruppo. Questi fattori casuali vengono introdotti nel modello e vanno ad affiancare gli effetti fissi nella parte lineare del modello stesso.

Chiamiamo y_{it} l'osservazione t del gruppo i . Diciamo quindi che t varia tra 1 e T_i , la numerosità del cluster i , che può variare di gruppo in gruppo. Alla stessa maniera indichiamo con \mathbf{x}_{it} un vettore contenente i valori dei predittori per l'osservazione t nel cluster i ; questo vettore si riferisce agli effetti fissi e moltiplicherà il vettore dei coefficienti per gli effetti fissi, indicato con $\boldsymbol{\beta}$. Indichiamo poi \mathbf{u}_i il vettore degli effetti casuali per il cluster i , comune a tutte le osservazioni appartenenti a questo gruppo e con \mathbf{z}_{it} un vettore di covariate per gli effetti casuali. In tal maniera la componente lineare di un modello GLMM assume la seguente forma compatta:

$$g(\mu_{it}) = \mathbf{x}_{it}^t \boldsymbol{\beta} + \mathbf{z}_{it}^t \mathbf{u}_i \quad (6.1)$$

dove g rappresenta come prima la funzione di link. Per la componente casuale \mathbf{u}_i si assume una distribuzione normale multivariata $N(\mathbf{0}, \Sigma)$.

Ciò significa che, al variare dei diversi gruppi considerati, ad alcuni coefficienti del modello è permesso di variare secondo una legge casuale (in questo caso normale). I gruppi considerati avranno quindi alcune caratteristiche simili fra loro, essenzialmente legate al fatto che essi provengono in un certo senso dalla stessa popolazione (che siano quindi confrontabili). Allo stesso tempo però si desidera mettere in evidenza la presenza di variabilità tra di essi e, sperabilmente, trarre delle conclusioni al riguardo. Per esempio, anticipando il discorso che faremo successivamente per i dati riguardanti le schede di dimissione per problemi di HCD negli ospedali lombardi presentati nel Capitolo 4, la presenza di un raggruppamento dei pazienti per ospedale può essere considerato come un effetto casuale: potremmo infatti aspettarci che diversi ospedali, a causa di fattori non misurabili (per esempio la preparazione dei suoi medici, il tipo di bacino di utenza della struttura, etc.), presentino delle proprie peculiari caratteristiche di cui tenere conto nel modello.

Si può quindi dire che modelli di questo tipo siano modelli a dua stadi: condizionatamente a \mathbf{u}_i , le osservazioni seguono un modello logit. Successivamente, le \mathbf{u}_i sono casuali. Ciò rende il procedimento di stima dei coefficienti più complesso rispetto ad un modello logit semplice. Non ci addentreremo

nelle metodologie numeriche per il problema della stima dei coefficienti per questi modelli. Molti dettagli e riferimenti sono contenuti in [1].

Come osservato prima, la variabilità degli \mathbf{u}_i induce una covarianza non negativa tra le risposte. Risposte appartenenti al medesimo gruppo tendono ad essere simili, quindi positivamente correlate, rispetto a risposte relative a soggetti appartenenti a gruppi differenti.

Nella Formula (6.1) la componente casuale entra nel modello allo stesso livello dei predittori lineari. Questa è una formulazione conveniente, soprattutto nel caso in cui il termine casuale serva a tenere conto di predittori omessi dal modello stesso perchè, ad esempio, non misurabili. La parte casuale è quindi utile a riflettere quei termini che avrebbero dovuto trovarsi nel modello se le relative variabili predittore fossero state incluse. I fattori casuali permettono quindi di tenere conto della over-dispersion per modelli lineari. Abbiamo già potuto osservare un meccanismo simile nel Capitolo 3.

6.2 Logistic-Normal Model

Introduciamo il modello cui siamo interessati, ovvero l'estensione del modello di regressione logistica con l'introduzione di effetti casuali, chiamato *logistic-normal model*. Vediamo il modello con effetto casuale normale univariato, che assume la seguente forma:

$$\text{logit}[\mathbb{P}(Y_{it} = 1|u_i)] = \mathbf{x}_{it}^t \boldsymbol{\beta} + u_i. \quad (6.2)$$

In questo caso quindi u_i sono variabili indipendenti e identicamente distribuite dalla legge $N(0, \sigma^2)$. In questo caso particolare, la componente casuale va ad impattare sul valore dell'intercetta del modello; per questo il modello può anche essere detto *random-intercept*.

Come già osservato nel Capitolo 5, modelli più generali possono utilizzare come link function l'inversa di una qualsiasi funzione cumulativa di probabilità, che possiamo indicare con Φ . Il modello logistico rappresenta il caso particolare in cui viene utilizzata la cumulativa inversa di una variabile logistica standard. Una valida alternativa potrebbe essere l'inversa di una normale standard; in tal caso, come già detto precedentemente nel Capitolo 5, il modello è detto probit.

Osserviamo anche che due variabili appartenenti al medesimo cluster, nonostante siano trattate come indipendenti condizionatamente a u_i , sono marginalmente correlate. Infatti, sfruttando la formula della varianza

condizionata, si ha:

$$\begin{aligned}\mathbb{C}ov(Y_{is}, Y_{it}) &= \mathbb{E}[\mathbb{C}ov(Y_{is}, Y_{it}|u_i)] + \mathbb{C}ov[\mathbb{E}(Y_{is}|u_i), \mathbb{E}(Y_{it}|u_i)] \\ &= 0 + \mathbb{C}ov[\Phi(\mathbf{x}_{is}^t \boldsymbol{\beta} + u_i), \Phi(\mathbf{x}_{it}^t \boldsymbol{\beta} + u_i)].\end{aligned}$$

Le funzioni contenute nella seconda covarianza sono termini monotoni crescenti in u_i ; di conseguenza la covarianza è non negativa.

Il focus dell'analisi di un modello GLMM è rivolto generalmente alla stima degli effetti fissi quando l'introduzione dell'effetto casuale ha lo scopo di modellizzare la correlazione tra osservazioni nel medesimo cluster. Spesso però la stima della variabilità dell'effetto casuale, che indichiamo con $\hat{\sigma}^2$, è utile per avere un stima del grado di eterogeneità della popolazione studiata.

6.2.1 Modello condizionato e modello marginale

Si può osservare come i parametri di un modello GLMM abbiano una doppia interpretazione, condizionale e marginale. Cominciamo dalla loro interpretazione condizionata all'effetto casuale.

Per un predittore che varia all'interno di un cluster, il relativo coefficiente si riferisce all'effetto *within-cluster* (all'interno del cluster) sulla risposta ad una variazione del predittore. Infatti all'interno del medesimo cluster, anche l'effetto casuale è fisso come gli altri predittori al variare del fattore di interesse. Il significato è quindi analogo a quello che assumono i coefficienti in un modello di regressione logistica senza effetti casuali. Per una variazione di una unità nella variabile di interesse, l'odds aumenta di un fattore e^β .

Se il predittore considerato assume sempre lo stesso valore all'interno di ogni cluster, il suo coefficiente si riferisce ad un effetto *between-cluster* (tra i cluster). La comparazione può essere svolta però solamente tra soggetti i cui effetti casuali assumano il medesimo valore, così come per gli altri predittori.

È possibile ricavare un modello marginale dal modello GLMM, in cui gli effetti vengono mediati su tutti i cluster. In tal maniera i coefficienti ottenuti non si riferiscono a comparazioni a valore fissato dell'effetto casuale. Ciò che si osserva è che gli effetti nel modello marginale, data la non linearità della link function, hanno un valore generalmente inferiore rispetto agli effetti specifici di ogni cluster per il GLMM. Vengono di conseguenza smorzati a causa della processo di media.

Nel modello introdotto nella Formula (6.1), $g(\mu_{it})$ rappresenta una media condizionale del tipo $\mathbb{E}(Y_{it}|u_i)$. Invertendo la link function e marginalizzando con la formula del valore atteso condizionato si ha:

$$\mathbb{E}(Y_{it}) = \mathbb{E}[\mathbb{E}(Y_{it}|\mathbf{u}_i)] = \int g^{-1}(\mathbf{x}_{it}^t \boldsymbol{\beta} + \mathbf{z}_{it}^t \mathbf{u}_i) f(\mathbf{u}_i; \Sigma) d\mathbf{u}_i,$$

dove $f(\mathbf{u}_i; \Sigma)$ è la funzione di densità di una $N(\mathbf{0}, \Sigma)$.

Se usiamo la link identità troviamo:

$$\mathbb{E}(Y_{it}) = \int (\mathbf{x}_{it}^t \boldsymbol{\beta} + \mathbf{z}_{it}^t \mathbf{u}_i) f(\mathbf{u}_i; \Sigma) d\mathbf{u}_i = \mathbf{x}_{it}^t \boldsymbol{\beta}.$$

Ciò non accade usando una link non lineare, quale ad esempio la funzione logit. In questo caso può risultare piuttosto complessa la soluzione dell'integrale, per cui esistono alcune relazioni approssimate. Ad esempio per il modello logistic-normal nel caso di valori piccoli di σ è stato mostrato che:

$$\mathbb{E}(Y_{it}) \approx \frac{\exp(c\mathbf{x}_{it}^t \boldsymbol{\beta})}{1 + \exp(c\mathbf{x}_{it}^t \boldsymbol{\beta})}$$

dove $c = (1 + 0.6\sigma^2)^{-1/2}$. Gli effetti nel modello marginale sono quindi equivalenti agli effetti nel modello condizionato moltiplicati per il coefficiente c . Di conseguenza gli effetti nel modello marginale risultano generalmente minori in valore assoluto. La discrepanza aumenta all'aumentare della deviazione standard σ .

6.2.2 Cenni sull'inferenza per i GLMM

Una volta fissato un modello GLMM, l'inferenza riguardo agli effetti fissi può essere eseguita nella medesima maniera del caso GLM.

Fare inferenza sugli effetti casuali invece può diventare più complesso. In alcuni casi ad esempio un modello può essere un caso particolare di un modello più complesso in cui una componente di varianza sia nulla. In questo caso non si applica la consueta inferenza sul rapporto di verosimiglianza.

La distribuzione asintotica della statistica basata sul rapporto di verosimiglianza è nota per la situazione più comune, ovvero quando si testa $H_0 : \sigma^2 = 0$ contro $H_1 : \sigma^2 > 0$ in un modello con un'unica componente di varianza. In questo caso la distribuzione nulla è una mistura equamente pesata tra una χ_0^2 , cioè una distribuzione degenera in 0, e una χ_1^2 . Se $\hat{\sigma} > 0$ e la statistica osservata vale t , il p-value per campioni grandi (cioè con molti cluster) vale $\frac{1}{2}\mathbb{P}(\chi_1^2 > t)$.

6.3 Un modello GLMM per i dati HCD

Torniamo ora ai dati relativi alle schede di dimissione ospedaliera di Regione Lombardia, e ripartiamo da dove ci siamo fermati al capitolo precedente.

Alla fine del Capitolo 5 abbiamo selezionato un modello che includeva gli effetti dell'età, del sesso, del numero di ricoveri subiti, dell'indice di salute

hidx ed anche gli effetti degli indicatori di SHOCK, PTCA, ICD e CABG. Vogliamo ora aggiungere a questo modello l'effetto del raggruppamento per ospedale, al fine di stimare la variabilità dell'effetto casuale, nonché di valutare l'entità degli effetti per gli ospedali considerati. Poniamo l'effetto casuale sull'intercetta; il nostro modello diventa quindi un modello di tipo *varying intercept*.

Il modello che consideriamo è quindi il seguente:

$$Y_{ij} \sim \text{Bernoulli}(p_{it})$$

$$\text{logit}(p_{it}) = \mathbf{x}_{it}^t \boldsymbol{\beta} + b_i.$$

Assumeremo che $b_i \sim N(0, \sigma^2)$. Per fittare il modello utilizzeremo la funzione `glmer()` presente nel pacchetto `lme4` di R.

```
glmm<-glmer(dead_intraH~ age + sex + hidx + npr
+ I(SHOCK!=0) + I(PTCA!=0) + I(ICD!=0)
+ I(CABG!=0) + (1 | hosp),
family=binomial, data=msm)
summary(glmm)
```

Otteniamo così il seguente output:

```
Generalized linear mixed model fit by the Laplace approximation
Formula: dead_intraH ~ age + sex + hidx + npr
          + I(SHOCK!=0) + I(PTCA!=0) + I(ICD!=0)
          + I(CABG != 0) + (1 | hosp)
Data: ms
AIC    BIC logLik deviance
23770 23852 -11875    23750
Random effects:
Groups Name          Variance Std.Dev.
hosp  (Intercept) 0.26323  0.51306
Number of obs: 26650, groups: hosp, 412

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.046677    0.151589 -33.29 < 2e-16 ***
age          0.039801    0.001799  22.13 < 2e-16 ***
sex1        -0.230538    0.034779  -6.63 3.39e-11 ***
hidx        0.086029    0.013534   6.36 2.06e-10 ***
```

```

npr                0.032503    0.006966    4.67 3.07e-06 ***
I(SHOCK != 0)TRUE  3.621756    0.097950    36.98 < 2e-16 ***
I(PTCA != 0)TRUE  -0.548024    0.084052    -6.52 7.03e-11 ***
I(ICD != 0)TRUE   -0.950011    0.125592    -7.56 3.90e-14 ***
I(CABG != 0)TRUE  0.755989    0.268191    2.82 0.00482 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

```

Correlation of Fixed Effects:

	(Intr)	age	sex1	hidx	npr	I(SH!0	I(PT!0	I(IC!0
age		-0.915						
sex1		0.086	-0.249					
hidx		-0.097	-0.107	0.084				
npr		-0.106	0.095	0.042	-0.563			
I(SHOCK!=0)		-0.081	0.061	-0.018	0.033	-0.012		
I(PTCA!=0)T		-0.080	0.060	0.053	0.059	-0.107	-0.079	
I(ICD!=0)TR		-0.099	0.085	0.063	0.053	-0.086	-0.030	-0.032
I(CABG!=0)T		-0.044	0.036	0.010	0.015	-0.004	-0.005	-0.003
								-0.006

Nella parte alta dell'output sono riportati alcuni indici di goodness of fit del modello. Possiamo vedere che l'AIC, pari a 23770, migliora rispetto al modello equivalente senza effetti casuali². Anche la deviance del modello diminuisce rispetto al precedente modello: abbiamo ora 23750 contro il precedente valore di 24112.98. Si ha una forte propensione per il modello in cui siano inclusi gli effetti casuali.

Scendendo troviamo la stima della varianza e della deviazione standard dell'effetto casuale, ovvero la stima $\hat{\sigma}^2$ di σ^2 . Abbiamo in particolare $\hat{\sigma}^2 = 0.26323$ che evidenzia una certa variabilità tra i diversi gruppi. Possiamo ad esempio testare la variazione in deviance tra questo modello e l'equivalente senza effetti casuali. Valutando questa differenza utilizzando i valori riportati poco sopra e sapendo, per quanto detto nei precedenti paragrafi, che sotto l'ipotesi $H_0 : \sigma^2 = 0$ questa statistica è una mistura di due χ^2 , otteniamo che il p-value, la cui espressione è $\frac{1}{2}\mathbb{P}(\chi_1^2 > 24100 - 23739)$, vale praticamente 0. Ci sentiamo quindi molto propensi a considerare significativa l'azione dell'effetto casuale.

Troviamo poi la tabella con la stima dei singoli coefficienti, il loro errore standard e la rispettiva significatività marginale. I risultati sono equivalenti a quelli per il modello senza effetti casuali.

Dal modello ottenuto possiamo estrarre le stime degli effetti dei singoli ospedali. In Figura 6.1 possiamo vedere alcuni grafici riassuntivi.

²L'AIC del precedente modello valeva 24130.98, circa 24131

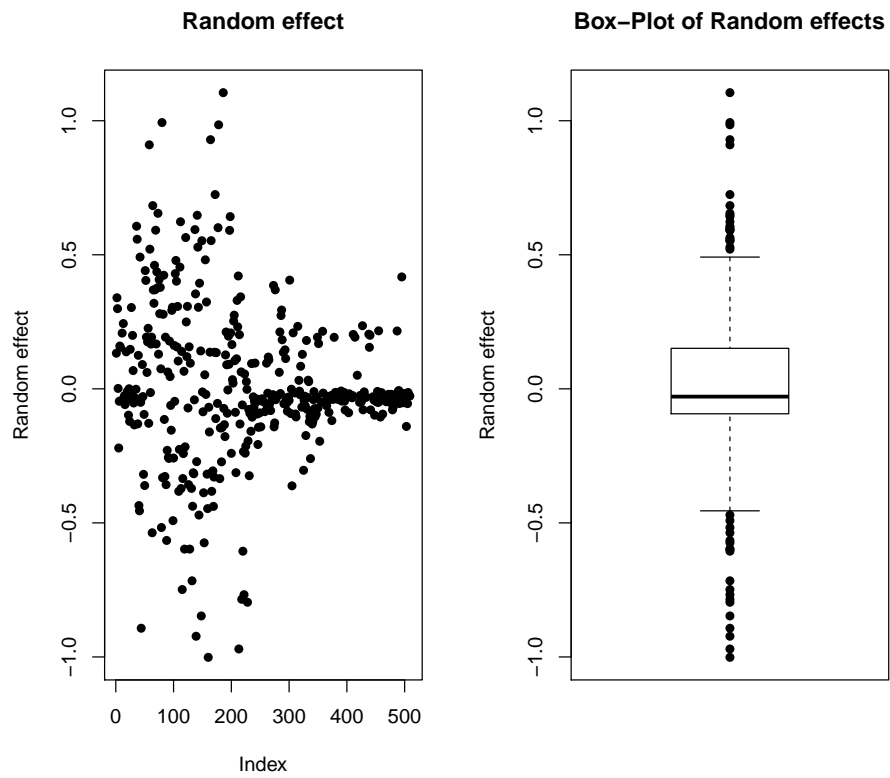


Figura 6.1: Box-Plot e Scatter-Plot per le stime degli effetti dei 412 ospedali considerati.

Osservando il grafico possiamo ancora una volta vedere come la variabilità dovuta al raggruppamento per ospedale sia tutt'altro che trascurabile. Infatti vi sono parecchi ospedali con effetti *estremi* rispetto alla popolazione degli ospedali osservati, come evidenziato dal box-plot in Figura 6.1 a destra.

Possiamo ricavare quali ospedali cadano tra gli outlier evidenziati dal box plot. I valori che delimitano i whiskers del grafico sono -0.45499698 e 0.49144505 ; i 22 ospedali che superano il baffo superiore sono quelli contraddistinti dai seguenti sodici: 36, 37, 58, 59, 64, 69, 73, 80, 112, 121, 137, 141, 142, 149, 164, 165, 172, 177, 178, 186, 197 e 198. I 20 che rimangono sotto il baffo inferiore sono invece i seguenti: 41, 44, 63, 79, 88, 99, 115, 119, 128, 132, 139, 144, 148, 153, 160, 213, 218, 220, 222 e 228.

Possiamo fare un'osservazione piuttosto interessante. Nel Capitolo 4 abbiamo riportato per i dati HCD diversi Funnel Plot. In particolare al paragrafo 4.2.1 abbiamo presentato il funnel plot senza correzioni per l'overdispersion e con limiti approssimati gaussiani. Andando a confrontare quali ospedali venivano segnalati come poco virtuosi, possiamo notare che alcuni ospedali vengono indicati anche come outlier dal box plot degli effetti casuali. In particolare sono gli ospedali 36, 37, 58, 64, 69, 73, 80, 141, 142, 164, 178, 197 e 198. Abbiamo quindi che ben 13 ospedali sui 22 indicati dal box plot venivano indicati anche dal precedente funnel plot. Per quelli contrassegnati da un effetto casuale troppo estremo inferiormente, abbiamo che gli ospedali 44, 119, 132, 148, 153, 160, 213, 218, 220, 222 e 228. Anche in questo caso 11 ospedali dei 20 segnalati venivano evidenziati anche nel funnel plot del Capitolo 4.

Tutto ciò è molto interessante. Vuol dire che vi sono alcune strutture che sono indicate come estreme attraverso diverse procedure di tipo inferenziale.

I Funnel Plot del Capitolo 4 però sono ancora piuttosto grezzi: essi non tengono conto delle covariate di interesse dei pazienti associati ai diversi ospedali. Sarà interessante nel successivo capitolo costruire dei nuovi grafici che tengano conto del modello fittato e quindi delle covariate di interesse. Vedremo poi di confrontare le conclusioni tratte dai grafici con le conclusioni ottenute dallo studio degli effetti casuali.

Capitolo 7

Funnel Plot e GLMM a confronto

GIUNTI a questo punto diamo uno sguardo a quanto è stato fin'ora presentato. Nei primi tre capitoli abbiamo presentato i Funnel Plot: abbiamo rimarcato la loro utilità e comodità nella presentazione visiva di risultati statistici, quali ad esempio la descrizione di performance di diversi soggetti; abbiamo sottolineato la loro efficacia rispetto ad altri strumenti riassuntivi, quali per esempio grafici di tipo caterpillar, e sottolineato come possano anche permettere lo studio di un fenomeno in maniera sequenziale. Abbiamo poi indagato le possibili correzioni per l'over-dispersion: se i dati sono più variabili di quanto il modello considerato prevederebbe, il rischio è quello di sottostimare la variabilità dell'indicatore di performance, segnalando come interessanti strutture che non lo sono. Nel Capitolo 4 abbiamo applicato le metodologie presentate ad un data set molto ampio, riguardante schede di dimissione ospedaliera per malattie di tipo cardiaco raccolte attraverso un progetto di monitoraggio di Regione Lombardia.

Nei Capitoli 5 e 6 abbiamo presentato gli elementi basilari dei modelli lineari generalizzati (ad effetti fissi e ad effetti misti), studiando e selezionando un modello per descrivere la probabilità di morte intra ospedaliera per il data set HCD. Il nostro proposito era quello di costruire un modello che prevedesse la probabilità di morte intra ospedaliera che tenesse conto delle caratteristiche di ogni paziente. Successivamente è stato aggiunto l'effetto del raggruppamento per ospedale. In particolare, dall'osservazione delle stime degli effetti casuali stimati attraverso il modello ad effetti misti finale costruito nel capitolo precedente, si è potuta notare una variabilità non trascurabile per gli effetti del raggruppamento per ospedale. Parecchi sono gli ospedali che presentano effetti molto ampi.

In questo capitolo vogliamo dapprima riconsiderare alcune tecniche per la costruzione di Funnel Plot presentate nei primi tre capitoli. In particolare vogliamo costruire dei nuovi Funnel Plot che tengano conto questa volta

delle covariate di interesse nella previsione della mortalità intra ospedaliera attraverso il modello GLM selezionato alla fine del Capitolo 5. Il nostro interesse sarà confrontare l'output di questi nuovi Plot con le strutture emergenti come 'notevoli' dall'osservazione degli effetti casuali provenienti dal modello GLMM.

Sottolineiamo che i Funnel Plot non verranno costruiti utilizzando direttamente il modello GLMM. In un tale modello viene già considerata l'over-dispersion dovuta alle differenze tra ospedali, includendola nel modello e non permettendo alle unità con effetti considerevoli di emergere alla nostra attenzione.

7.1 Indici di performance, di precisione e limiti di controllo

Dedichiamoci per prima cosa alla specificazione dei nuovi Funnel Plot. Come già detto nel Capitolo 1, per costruire un Funnel Plot è necessario selezionare quattro elementi base:

1. Un indicatore Y ;
2. Un target θ_0 ;
3. Un parametro di precisione ρ ;
4. La forma dei limiti di controllo.

Per costruire questa nuova serie di Funnel Plot utilizziamo la procedura già indicata alla sezione 1.2.3. Richiamiamo brevemente: in ogni struttura sono osservati O (*observed*) eventi, mentre il modello prevede che si verifichino E (*expected*) eventi. Per i nostri dati, O equivale al numero di morti osservate tra i pazienti associati ad un certo ospedale, mentre E , l'*expectation*, è il numero di morti attese in un certo ospedale. Valutare E per ogni struttura è chiaramente il punto più delicato di questa procedura; disponendo di un modello questo passo è ora semplice: infatti il modello GLM ci permette di prevedere per ogni paziente la probabilità di morte intra ospedaliera; di conseguenza la somma delle probabilità per i pazienti associati ad un certo ospedale corrisponde proprio ad E per l'ospedale in questione, cioè il valore atteso delle morti. Il nostro indicatore Y sarà quindi l'SMR (*Standardized Mortality Ratio*), il rapporto O/E .

Per una struttura sotto controllo considereremo (per il modello senza aggiustamenti per l'over-dispersion) che O sia distribuito come una $Poisson(E)$.

Per quanto detto alla sezione 1.2.3, il valore di target θ_0 vale di conseguenza 1. Tanto più si ha un numero di morti osservate O differente dal numero di morti previste dal modello GLM E , tanto più la struttura evidenzia un comportamento fuori controllo.

Sapendo poi che la varianza di una variabile di Poisson di parametro E vale proprio E , abbiamo che la varianza di SMR vale $1/E$. Una misura di precisione è quindi l'*expectation*.

Da ultimo, considereremo l'approssimazione con limiti Gaussiani. L'espressione chiaramente dipende da quale delle metodologie viene scelta tra quelle che abbiamo presentato nei primi capitoli.

7.2 Funnel Plot senza correzioni

Considereremo i limiti di precisione approssimati $y_p = 1 + z_p \sqrt{1/E}$, come indicati già nella sezione 1.2.3. Fissiamo i valori di p a 0.025, 0.975, 0.001 e 0.999; in tal maniera avremo delle bande al 95% e al 99.8%.

In Figura 7.1 riportiamo quindi il Funnel Plot per l'SMR contro il numero atteso di morti. Sono stati riportati solo gli ospedali con almeno 5 casi. Dal grafico emergono 10 strutture con SMR significativamente più alto di 1, ovvero gli ospedali contraddistinti dai codici 13 (non compare sul grafico in quanto l'SMR relativo è fuori scala), 36, 58, 64, 80, 164, 178, 186, 197 e 198. Ad essi si aggiungono le strutture 37, 73, 112, 121, 141, 172, 177, 427, 455 e 495 se consideriamo anche quelle che cadono nella banda superiore di warning. Vengono segnalate anche 6 strutture virtuose: gli ospedali 132, 139, 160, 169, 213 e 222. Se consideriamo anche la banda di warning inferiore, segnaliamo anche le strutture contrassegnate dai codici 40, 44, 48, 88, 113, 115, 119, 148, 153, 159, 218, 220, 228 e 231. Sottolineiamo che di questi, 5 ospedali sono caratterizzati da SMR nullo.

Confrontiamo le indicazioni di questo metodo con quanto emerge dal modello GLMM.

Osserviamo la Figura 7.2. Nel grafico a sinistra abbiamo lo scatter plot delle stime dei random effect per i 412 ospedali studiati col modello GLMM. Sono stati evidenziati quegli ospedali che cadono oltre i whiskers del relativo box plot (vedi Figura 6.1, pannello a destra). In particolare i più estremi, quelli che cadono oltre i quantili al 2.5% e al 97.5% sono colorati diversamente:

- in rosso quelli che superano il quantile al 97.5%;
- in blu quelli sotto il quantile al 2.5%.

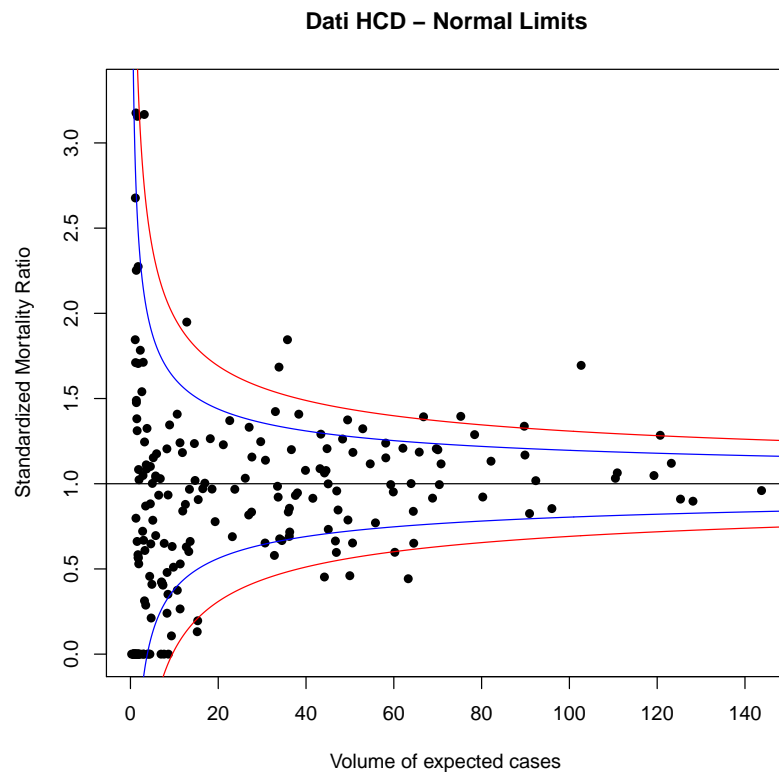


Figura 7.1: Funnel Plot per i dati HCD.

Nel grafico a destra vediamo il Funnel Plot come costruito in Figura 7.1, sul quale sono stati riportati gli ospedali evidenziati nel pannello a sinistra. In tal maniera possiamo vedere dove si collocano le strutture indicate come sospette dall'osservazione degli effetti casuali. Possiamo notare come già su questo grafico che non ha subito correzioni per eventuali over-dispersion la grande maggioranza delle strutture con effetti casuali molto grandi o molto piccoli venga anche evidenziata attraverso il Funnel Plot basato sul modello GLM. Possiamo notare che le strutture evidenziate di rosso effettivamente giacciono tutte oltre il limite al 97.5%; anche alcune strutture evidenziate in giallo giacciono oltre questo limite. A meno delle strutture con pochissime morti previste, non vi sono strutture non evidenziate nei pressi dei limiti di confidenza. Lo stesso discorso vale per le strutture evidenziate in blu: anch'esse cadono tutte sotto il limite di confidenza al 2.5%. Alcune strutture evidenziate in verde cadono anch'esse sotto questo limite. Osserviamo però che alcune strutture non segnalate dall'osservazione degli effetti casuali sono invece segnalate dal Funnel Plot nella fascia di warning inferiore.

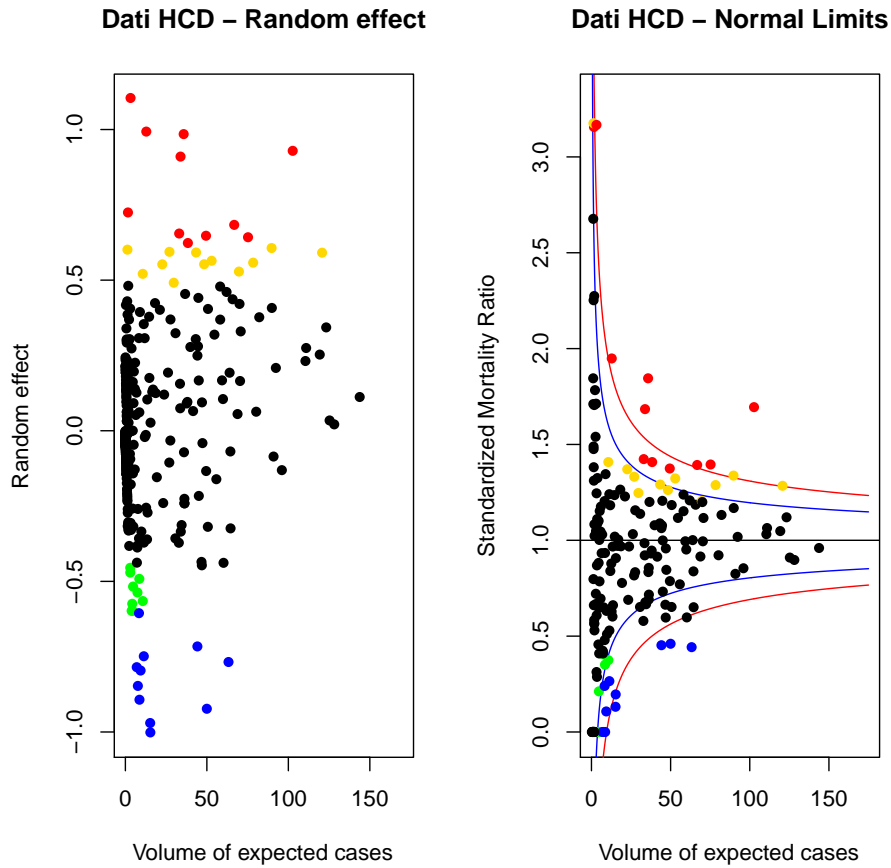


Figura 7.2: A sinistra lo scatter plot degli effetti casuali stimati dal modello GLMM, a destra il Funnel Plot per l'SMR, entrambi plottati contro le morti attese. Su entrambi i grafici sono evidenziati gli outlier indicati dal box plot degli effetti casuali. In particolare in rosso gli effetti oltre il quantile al 97.5% e in blu gli effetti sotto il quantile al 2.5%.

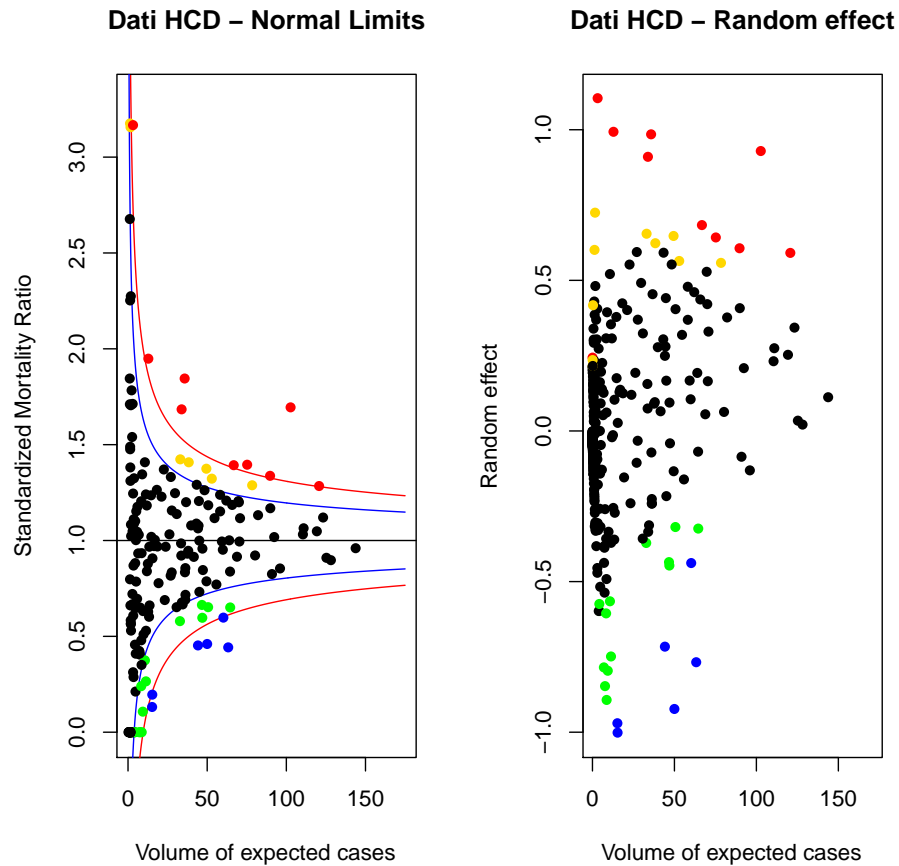


Figura 7.3: A sinistra il Funnel Plot per l'SMR contro le morti attese, a destra lo scatter plot per gli effetti casuali stimati dal modello GLMM. Su entrambi i grafici sono evidenziati gli ospedali indicati dal Funnel Plot. In particolare in rosso gli effetti oltre il limite al 99.9% e in blu gli effetti sotto il limite al 0.1%.

In Figura 7.3 possiamo osservare il confronto inverso; nel pannello di sinistra evidenziamo con diversi colori gli ospedali che cadono nelle fasce di warning e quelli che cadono oltre i limiti più esterni. Le medesime strutture sono evidenziate (con gli stessi colori) nel grafico di destra. Per quanto riguarda le strutture con SMR alto, possiamo vedere come esse siano anche caratterizzate da effetti casuali alti: sono poche le strutture con effetto casuale stimato maggiore di 0.49145^1 che non siano evidenziate anche dal Funnel Plot. Le strutture evidenziate con colore verde e blu sono invece più sparse a causa di quelle strutture nere che cadono oltre i limiti di controllo in Figura 7.2. Possiamo però osservare che tutte le strutture messe in evidenza sul Funnel Plot cadono fuori dalla zona in cui si accumulano la maggior parte degli effetti casuali.

7.3 Funnel Plot corretti per l'over-dispersion

Nei grafici appena mostrati abbiamo potuto notare un aspetto che ci si poteva attendere di osservare: che le strutture con effetti casuali più estremi tendono contemporaneamente a posizionarsi nelle zone più estreme del Funnel Plot. Abbiamo sottolineato che il mixing è buono ad un primo sguardo. Vorremmo però vedere se è possibile fare di meglio. Infatti, soprattutto nella fascia inferiore, il Funnel Plot segnala strutture con effetti casuali non notevoli. Proviamo quindi ad introdurre delle correzioni in cui i limiti vengano ampliati.

7.3.1 Correzione moltiplicativa

Utilizziamo la correzione moltiplicativa presentata nel Capitolo 2.

Ricordiamo che con questo approccio la varianza nulla viene accresciuta con un coefficiente moltiplicativo ϕ in modo tale che i limiti di confidenza diventino $y_p = \theta_0 + z_p \sqrt{\phi g(\theta_0)/\rho}$. Nel nostro caso abbiamo $y_p = 1 + z_p \sqrt{\phi/E}$. La stima di ϕ veniva ottenuta dalla media dei quadrati dei residui di Pearson (calcolati sotto l'ipotesi di assenza di over-dispersion). Nel nostro caso:

$$z_i = \frac{y_i - 1}{\sqrt{1/E}}$$

dove y_i rappresenta il rapporto osservati-attesi per la struttura i , e 1 è il valore di target θ_0 .

Nel Capitolo 2 abbiamo anche fatto notare che spesso è buona norma utilizzare una stima di tipo winsorized per ϕ , al fine di non sovrastimare

¹Ovvero la soglia del whisker superiore nel box plot dei random effect.

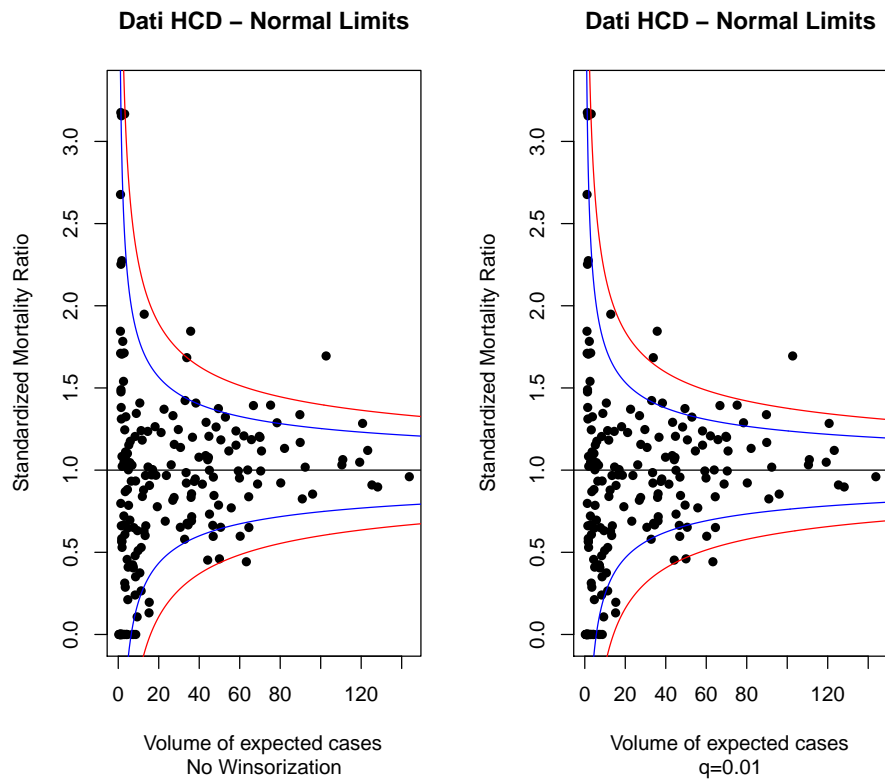


Figura 7.4: Funnel Plot per i dati HCD corretti con il coefficiente moltiplicativo ϕ . A sinistra il grafico senza winsorization, a destra il grafico con Winsorization pari a $q = 0.01$.

il coefficiente moltiplicativo, rendendo contemporaneamente la stima di ϕ robusta rispetto ad eventuali unità con residui di Pearson molto grandi e quindi fuori anche dal modello corretto. Per la stima winsorized di ϕ è necessario fissare un valore di q , cioè il livello di winsorization desiderato. Rimarchiamo che non vi sono regole per fissarne il valore, ma è necessario regolarlo secondo l'esperienza o a seconda del problema affrontato.

Passando ai dati relativi alla morte intra ospedaliera negli ospedali lombardi, useremo la stima di ϕ senza winsorization e con due livelli di winsorization: $q = 0.01$ e $q = 0.05$. Nel primo caso otteniamo una stima $\hat{\phi} = 1.667902$, nel secondo caso $\hat{\phi} = 1.483632$ e nel terzo caso $\hat{\phi} = 1.05764$. Il valore di soglia suggerito nel Capitolo 2, nonché in [13], oltre il quale assumere l'overdispersion vale $1 + 2\sqrt{\frac{2}{412}} = 1.139347$. Quindi riportiamo in Figura 7.4 i Funnel Plot corretti relativi ai primi due casi. Nel caso $q = 0.05$ è possibile non considerare la correzione, che avrebbe un impatto minimo.

Notiamo che la differenza tra i due grafici di Figura 7.4 è piuttosto contenuta. Un minimo livello di winsorization in generale è comunque utile e consigliato. Anche rispetto al grafico di Figura 7.1, le differenze sono contenute e le indicazioni che ci vengono date sono abbastanza simili. Osservando questi grafici quindi si potrebbe anche pensare di non considerare l'effetto dell'over-dipersion particolarmente incisivo.

Facciamo ora il passaggio successivo e osserviamo quali delle strutture che emergono dall'osservazione dei random effect vengono anche indicate dai Funnel Plot. In Figura 7.5 vediamo i medesimi grafici di Figura 7.4 nei quali sono stati colorati i punti relativi agli ospedali indicati come outliers nel box plot dei random effect. In rosso sono indicati gli ospedali il cui effetto casuale supera il quantile al 97.5% e in blu quelli con effetto casuale sotto il quantile al 2.5%, equivalentemente a quanto fatto prima.

Per quanto riguarda la fascia superiore, i limiti al 97.5% mettono bene in evidenza le strutture rosse, che sono appunto quelle che superano il quantile al 97.5% degli effetti casuali. Lo stesso avviene nella fascia inferiore, dove le strutture blu, cioè quelle il cui effetto rimane sotto il quantile al 2.5%, sono tutte evidenziate dal limite al 2.5%. Le altre strutture evidenziate rimangono abbastanza adese al limite blu. Grazie al lieve allargamento dei limiti, diminuiscono le strutture segnalate dai Funnel Plot nella fascia di warning inferiore che siano però caratterizzate da stime contenute del random effect.

In Figura 7.6 possiamo vedere dove si posizionano sullo scatter plot dei random effect stimati le strutture indicate dai Funnel Plot. La situazione è simile a quella riportata in Figura 7.3.

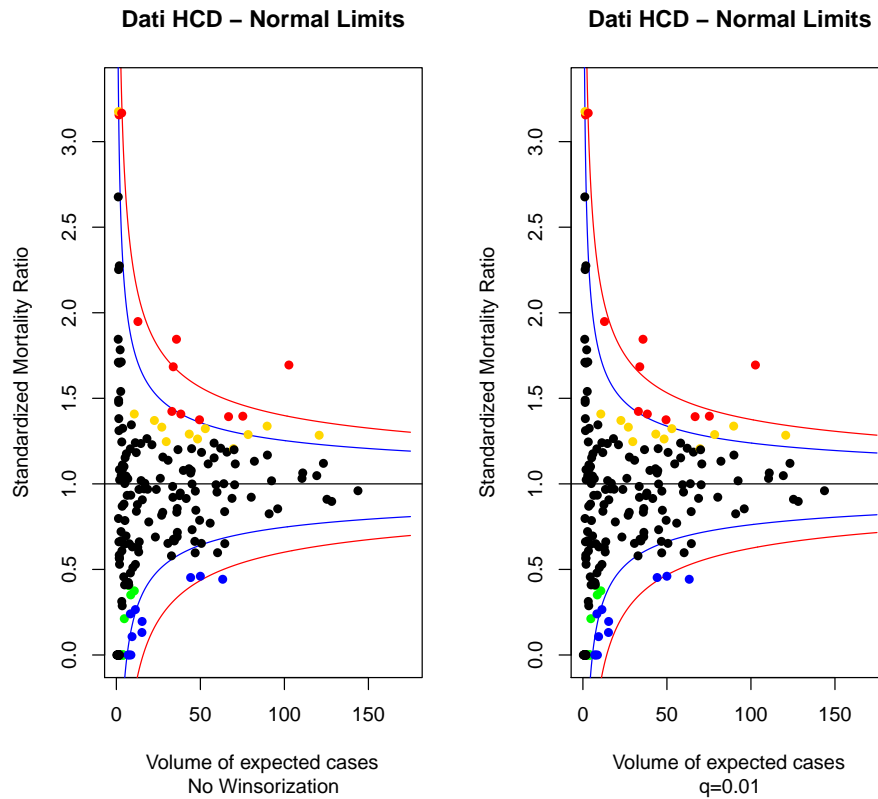


Figura 7.5: Funnel Plot per i dati HCD con correzione moltiplicativa. Sono evidenziati gli ospedali con effetti casuali stimati estremi. In rosso quelli oltre il quantile al 97.5%, in blu quelli sotto il quantile al 2.5%

7.3.2 Correzione additiva

Sfruttiamo adesso il secondo approccio proposto nel 2: in questo caso la varianza nulla è accresciuta con un fattore additivo τ^2 .

Ricordiamo che una volta stimato il valore di ϕ come già indicato per la correzione di tipo moltiplicativo, il valore di τ^2 si può stimare con la formula (2.4) come segue:

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I - 1)}{\sum_i w_i - \sum_i w_i^2 / \sum w_i}.$$

Ricordiamo anche che $w_i = 1/\sigma_i^2$. Nel nostro caso abbiamo che $w_i = E_i$. Il fattore τ^2 viene poi aggiunto direttamente alla varianza nulla del modello. In tal maniera i nostri limiti normali diventano $y_p = 1 + z_p \sqrt{1/E + \tau^2}$.

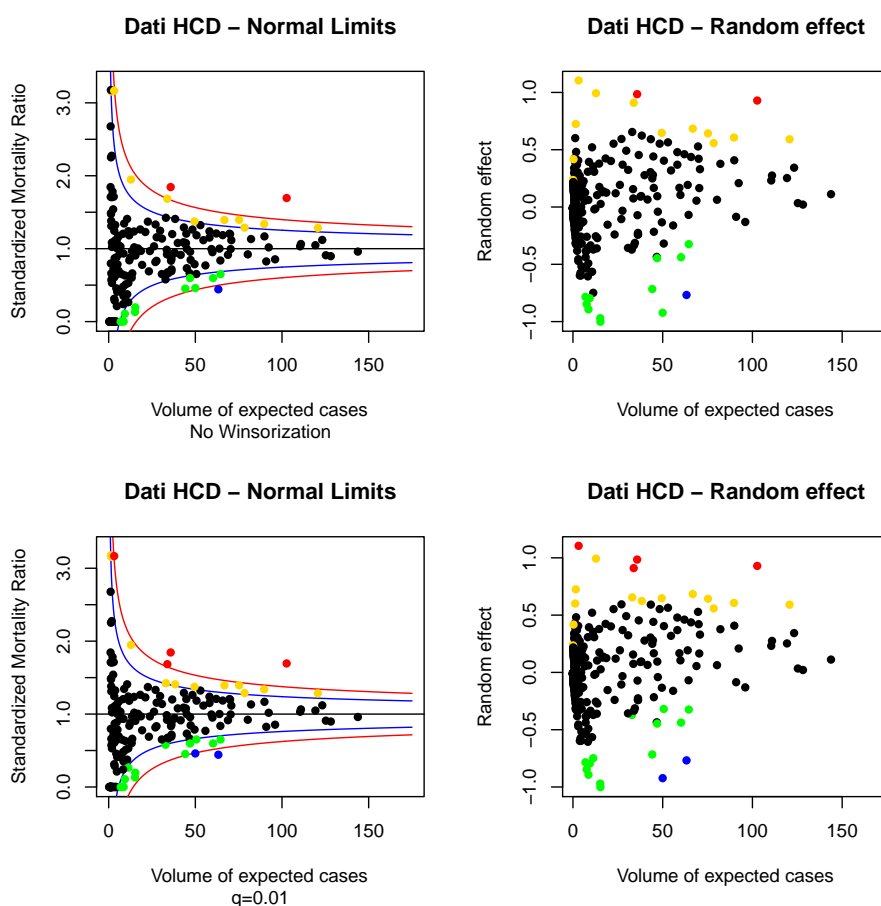


Figura 7.6: A sinistra i Funnel Plot per l'SMR contro le morti attese con correzione moltiplicativa, a destra gli scatter plot per gli effetti casuali stimati dal modello GLMM. Su tutti i grafici sono evidenziati gli ospedali indicati dal Funnel Plot. In particolare in rosso gli ospedali oltre il limite al 99.9% e in blu gli effetti sotto il limite al 0.1%.

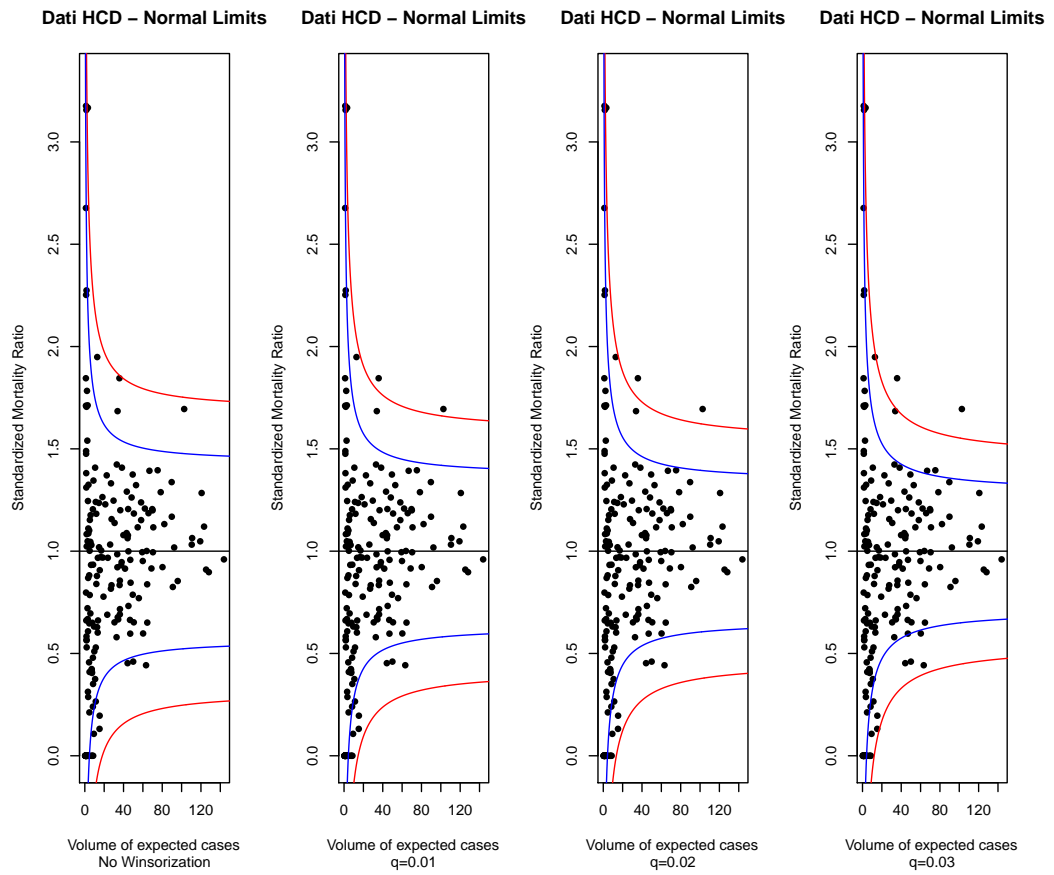


Figura 7.7: Funnel Plot per i dati HCD corretti con il coefficiente additivo τ^2 .

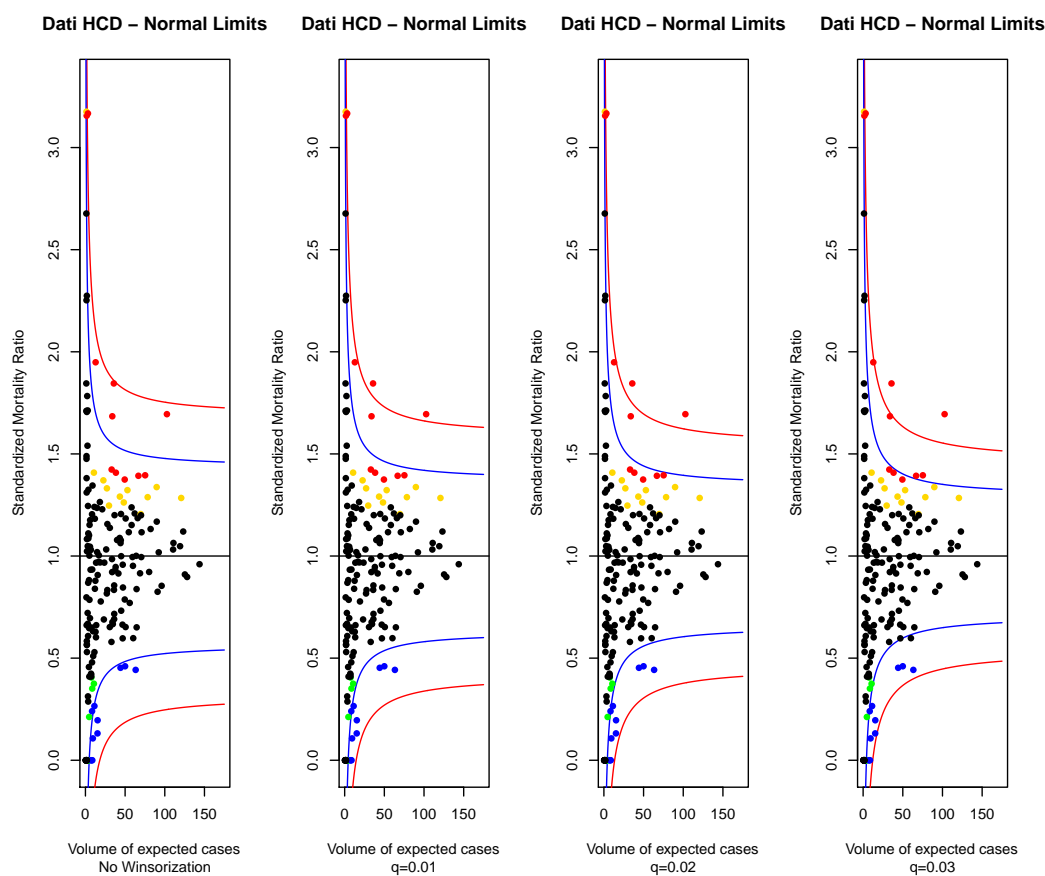


Figura 7.8: Funnel Plot per i dati HCD con correzione additiva. Sono evidenziati gli ospedali con effetti casuali estremi. In rosso quelli oltre il quantile al 97.5%, in blu quelli sotto il quantile al 2.5%

In Figura 7.7 riportiamo i Funnel Plot corretti additivamente per quattro possibili valori di winsorization: 0, 0.01, 0.02 e 0.03. Come già osservato nel Capitolo 2, maggiore è la winsorization, più contenuto è il valore di ϕ e di conseguenza il valore di τ^2 . Perciò i limiti sono più stretti e più vicini ai limiti costruiti senza correzione. I valori di τ^2 ottenuti nei quattro casi sono: 0.04942414, 0.03583769, 0.03064935 e 0.02205671. Come prima siamo interessati a vedere quali strutture sono evidenziate dai Funnel Plot in relazione ai loro effetti casuali.

In Figura 7.8 riportiamo i Funnel Plot di Figura 7.7 su cui sono evidenziati gli ospedali con effetti casuali stimati estremi. I colori sono equivalenti a quelli già utilizzati nelle precedenti sezioni. Possiamo asserire che, tra i livelli di winsorization utilizzati, gli ultimi due ($q = 0.02$ e $q = 0.03$) sono

quelli con prestazioni migliori. Vediamo che quasi tutte le strutture rosse (quelle oltre il quantile al 97.5%) cadono oltre il limite di warning superiore (appunto il limite al 97.5%), così come tutte le strutture evidenziate in blu (quelle sotto il quantile al 2.5%) cadono sotto il limite inferiore di warning (appunto il limite al 2.5%). Livelli di winsorization più bassa evidenziano alcune strutture estreme che non vengono segnalate dal grafico. Tutto ciò è chiaramente molto interessante, anche confrontando con i grafici fino ad ora ottenuti. In alcuni dei grafici osservati precedentemente ad esempio accadeva che alcune strutture evidenziate in giallo cadessero oltre il limite superiore di warning, oppure che strutture non evidenziate dal modello GLMM cadessero al di sotto del limite inferiore di warning. Ciò non accade in quest'ultima serie di grafici.

Vediamo ora in Figura 7.9 il confronto inverso e vediamo dove si posizionano gli effetti casuali delle strutture indicate dai Funnel Plot. Scegliamo come valori di winsorization solo i due valori più promettenti: $q = 0.02$ e $q = 0.03$.

Possiamo vedere innanzi tutto che le strutture colorate di verde rimangono molto compatte sullo scatter plot degli effetti casuali. Osservando le figure equivalenti delle precedenti sezioni, si può notare che le strutture della fascia inferiore erano molto più diffuse, mischiandosi con strutture non evidenziate. Osserviamo che non compaiono strutture evidenziate in blu. Per quanto riguarda le strutture con SMR più alto, osserviamo che in questo caso vi sono alcuni ospedali (uno segnalato in rosso e tre segnalati in giallo) che sul grafico degli effetti casuali si distaccano dal gruppo principale.

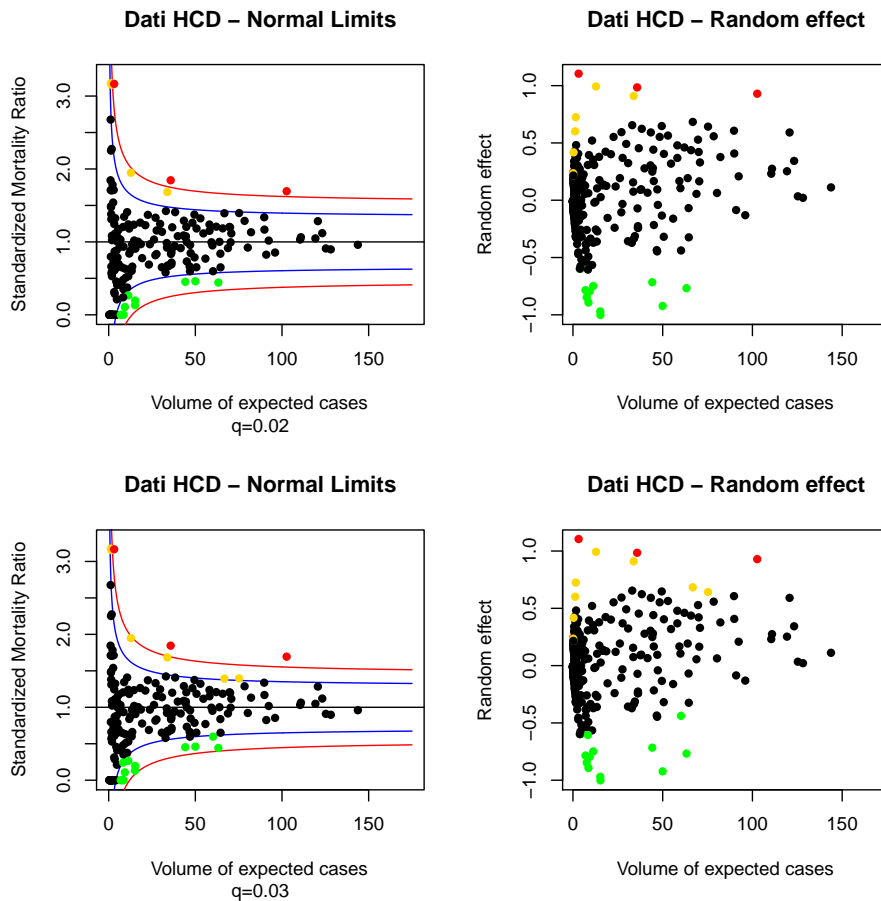


Figura 7.9: A sinistra i Funnel Plot per l'SMR contro le morti attese con correzione additiva, a destra gli scatter plot per gli effetti casuali stimati dal modello GLMM. Su tutti i grafici sono evidenziati gli ospedali indicati dal Funnel Plot. In particolare in rosso gli ospedali oltre il limite al 99.9%. Non vi sono strutture sotto il limite allo 0.1%.

7.4 Funnel Plot e modello *random effect*

Proseguiamo costruendo in questa sezione i tre tipi di Funnel Plot consigliati nel Capitolo 3, in cui abbiamo considerato un semplice modello random effect e abbiamo suggerito tre strategie per la ricerca di strutture interessanti. Ricordiamo che si può usare questo modello per cercare le seguenti tre categorie di strutture:

1. Outlier del modello *common mean*; in questo approccio viene considerata nulla la varianza aggiuntiva dovuta alla differenza tra le strutture, che vengono considerate provenienti dalla medesima distribuzione. Il Funnel Plot che si ottiene è equivalente a quello costruito senza correzioni.
2. Outlier del modello *random effect*; in questo approccio cerchiamo quelle unità che presumibilmente hanno una media μ_i per la variabile di controllo Y diversa dalla media comune che indichiamo con μ . Il Funnel Plot è equivalente a quello corretto col fattore additivo.
3. Estremi del modello *random effect*; in questo caso siamo invece interessati a quelle unità le cui realizzazioni della variabile di controllo Y siano troppo lontane dalla media comune μ ; quelle strutture quindi che, anche se giustificate dal modello, hanno presentato valori inusuali rispetto alla media μ .

Per quanto riguarda il terzo approccio, l'espressione dei limiti di controllo può essere ricavata direttamente dalla formula (3.6), scritta nel Capitolo 3 per una situazione identica:

$$y_p = 1 + z_p \sqrt{\frac{1 + E\tau^2}{E^2\tau^2}}.$$

La quantità τ^2 viene stimata come già indicato nella precedente sezione (formula (2.4)).

Usiamo i medesimi livelli di winsorization già usati in precedenza e riportiamo nelle Figure 7.10, 7.11, 7.12 e 7.13 i risultati.

Per ogni tripletta di grafici, il primo a sinistra è sempre equivalente al Funnel Plot costruito col modello nullo, quindi senza alcuna correzione; il grafico centrale equivale invece al Funnel Plot con correzione additiva. Osserviamo poi che, all'aumentare della winsorization e quindi al diminuire della stima di ϕ (quindi anche della stima di τ^2), i limiti riportati nel grafico centrale si restringono progressivamente. Al contempo i limiti riportati nel terzo grafico tendono ad ampliarsi. Questo effetto era dovuto al fatto che, coll'approccio

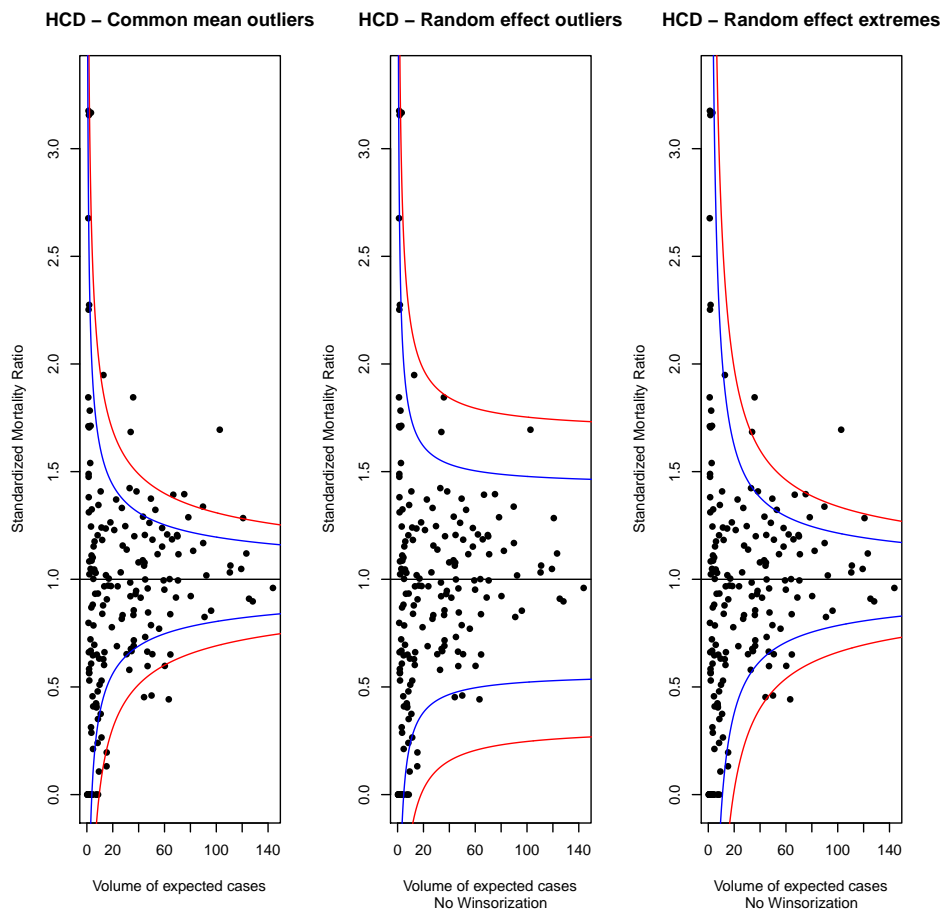


Figura 7.10: Funnel Plot basati sul modello *random effect*. $q = 0$, $\tau^2 = 0.049$.

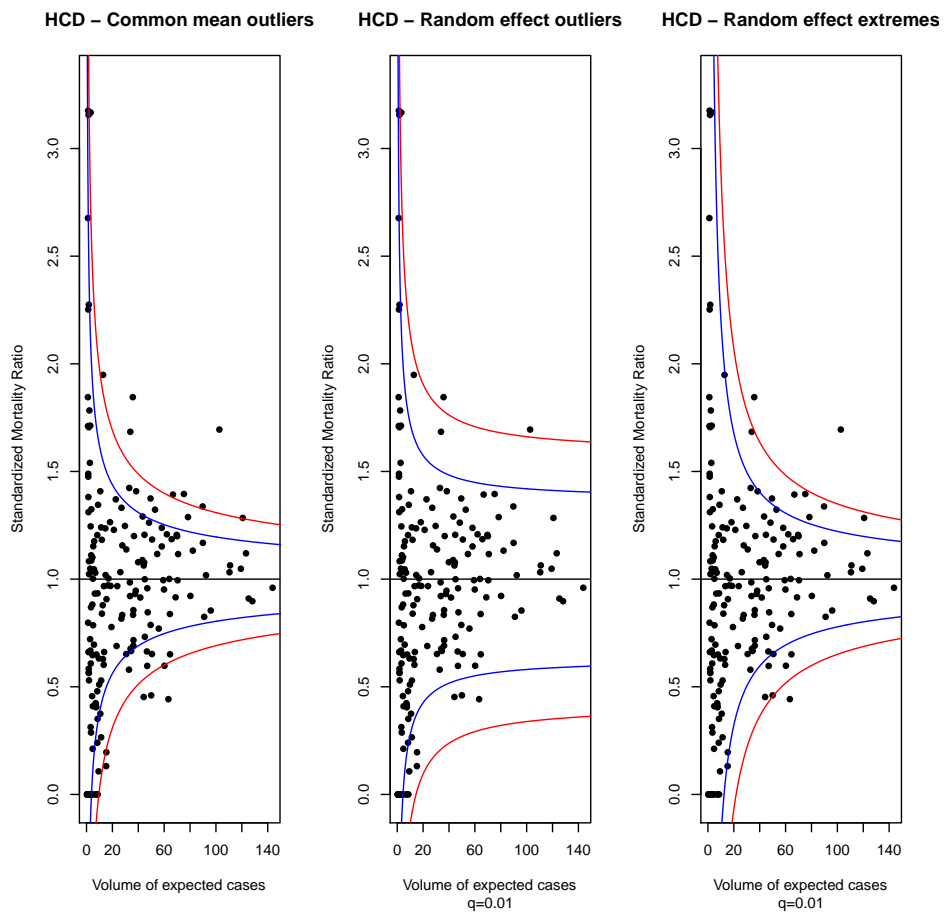


Figura 7.11: Funnel Plot basati sul modello *random effect*. $q = 0.01$, $\tau^2 = 0.036$.

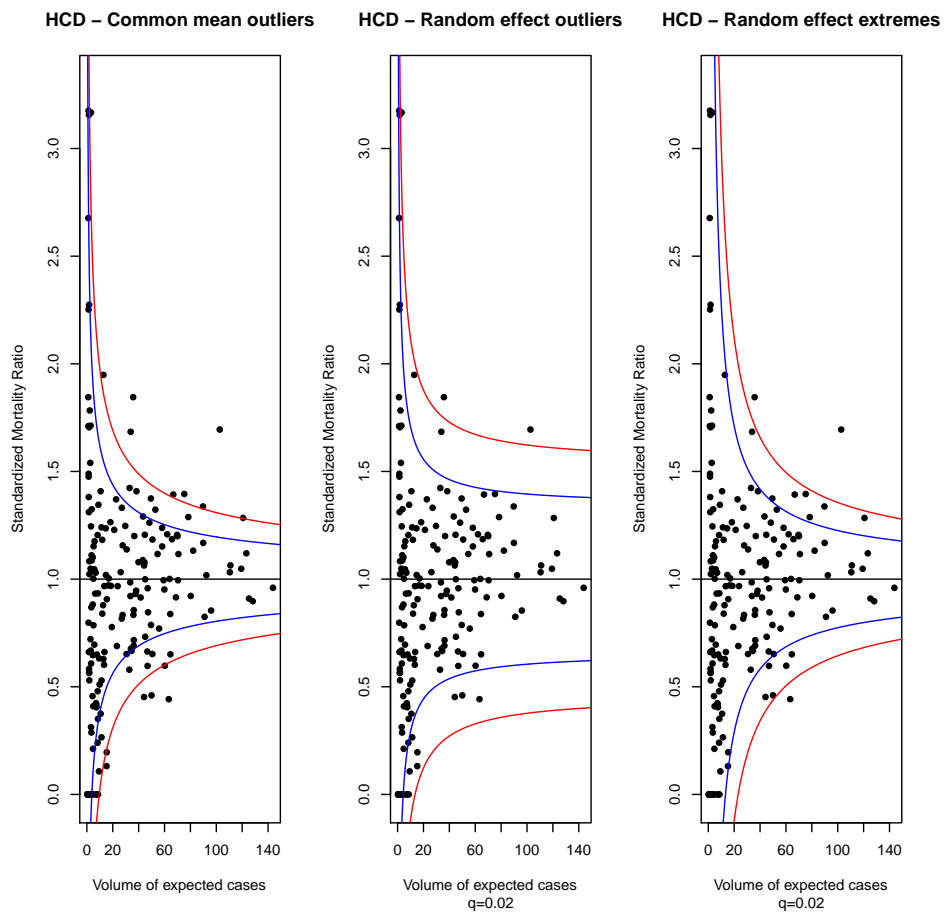


Figura 7.12: Funnel Plot basati sul modello *random effect*. $q = 0.02$, $\tau^2 = 0.030$.

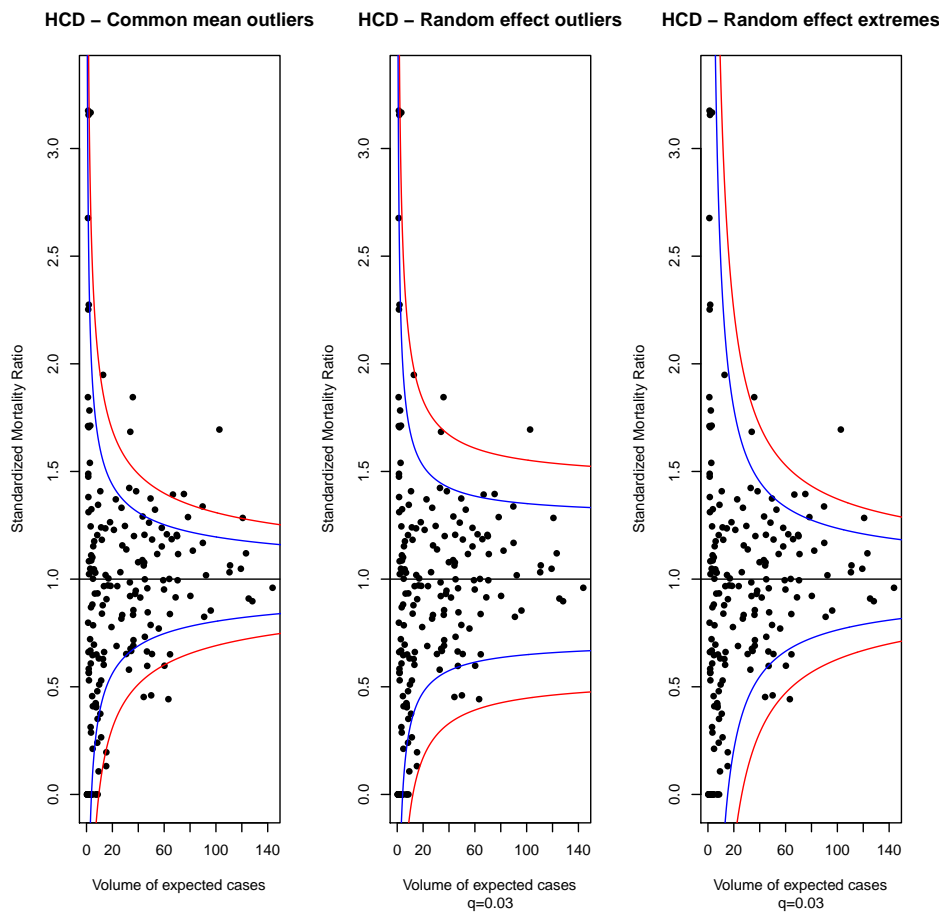


Figura 7.13: Funnel Plot basati sul modello *random effect*. $q = 0.03$, $\tau^2 = 0.022$.

per la ricerca degli outlier dal modello, si è interessati ad individuare quelle strutture la cui media μ_i sia molto grande rispetto alla variabilità delle stesse. Un valore più piccolo della stima di τ^2 implica una minore variabilità tra le medie delle diverse unità e quindi limiti più stretti. Nel terzo approccio ci si concentra invece sulle realizzazioni dell'indicatore di interesse² Y . Un valore piccolo di τ^2 implica poca variabilità tra le medie delle diverse strutture e di conseguenza gli estremi del modello saranno ricercati dal terzo approccio più lontano dalla media comune, cioè dal valore di target; nel grafico a destra si hanno quindi limiti tendenti ad allargarsi. Queste considerazioni possono trovarsi in maniera più estesa nel Capitolo 3 e in [6]. L'effetto è comunque contenuto dato che le stime di τ^2 non sono molto grandi.

Ora andiamo a controllare come si collocano i limiti rispetto alle strutture con effetti casuali notevoli. Tracciamo i grafici per i livelli di winsorization $q = 0.02$ e $q = 0.03$.

Confrontando le due triplete di grafici delle Figure 7.14 e 7.15, possiamo vedere che le indicazioni dei grafici a sinistra sono abbastanza simili a quelle dei grafici a destra (quelli per la ricerca degli estremi), maggiormente nella prima Figura. Nella seconda Figura si può notare che i limiti di confidenza per il grafico di ricerca degli estremi sono effettivamente più larghi. In questi grafici accade però che non solo le strutture evidenziate in rosso cadano oltre il limite al 97.5%, ma anche alcune unità evidenziate in giallo. Allo stesso modo sotto il limite al 2.5% cadono anche strutture non evidenziate. L'eventuale utilizzo della terza metodologia quindi non migliora le prestazioni del Funnel Plot.

Tra i tre grafici costruibili con questi approcci quello centrale resta il più promettente: con entrambe le winsorizations riesce ad evidenziare in modo corretto praticamente tutte le strutture evidenziate in blu (con $q = 0.03$ verrebbe indicata anche una struttura non evidenziata, comunque abbastanza adesa al limite di confidenza), e quasi tutte le strutture rosse: di queste se ne possono notare tre che non vengono segnalate dal limite al 97.5%, ma che cadono comunque molto vicine alla zona di warning. Queste osservazioni sono state già fatte nella sezione 7.3.2.

Nelle Figure 7.16 e 7.17 possiamo vedere i confronti inversi.

Si può facilmente notare come le unità evidenziate dai Funnel Plot per la ricerca degli outliers sul grafico degli effetti casuali dal modello *random effects* (grafici al centro) siano molto più compatte rispetto alle unità indicate dai Funnel Plot per la ricerca degli estremi (grafici a destra) e degli outliers dal modello common mean (grafici a sinistra). In particolare le prestazioni peggiori sembrerebbero proprio quelle date dai grafici di destra, nei quali le

²Lo Standardized Mortality Rate nel nostro caso.

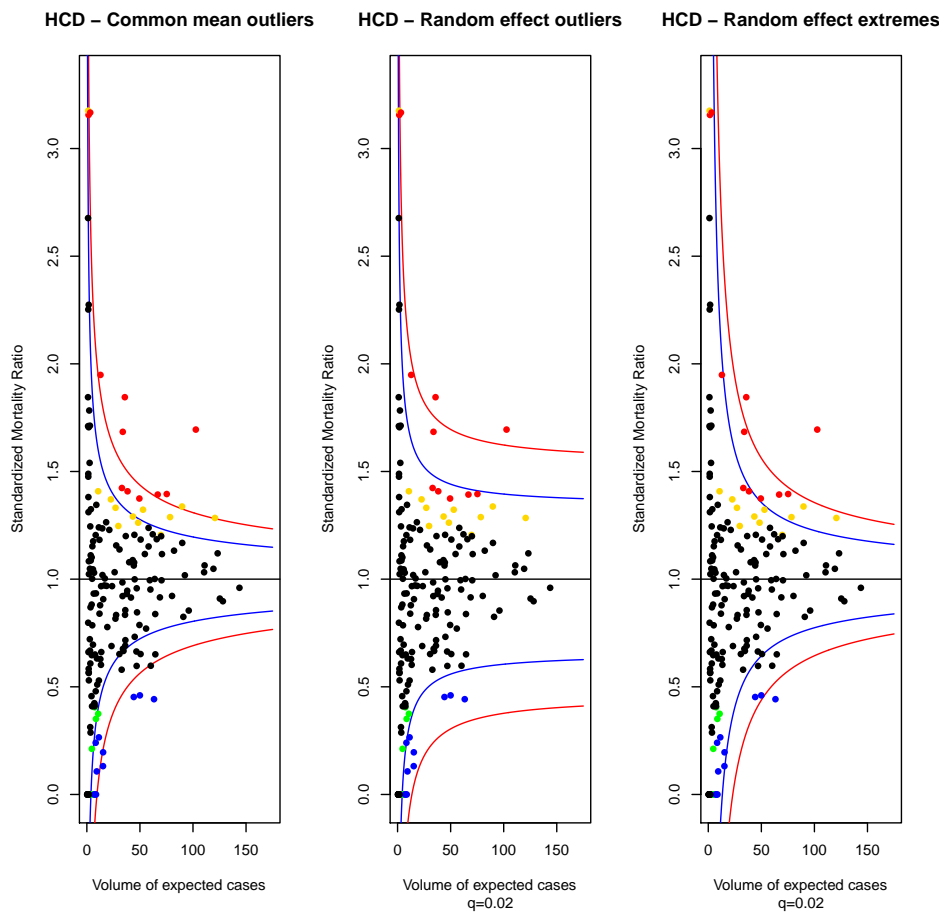


Figura 7.14: Funnel Plot basati sul modello *random effect*, con evidenziate le strutture affette da effetti casuali notevoli ($q = 0.02$).

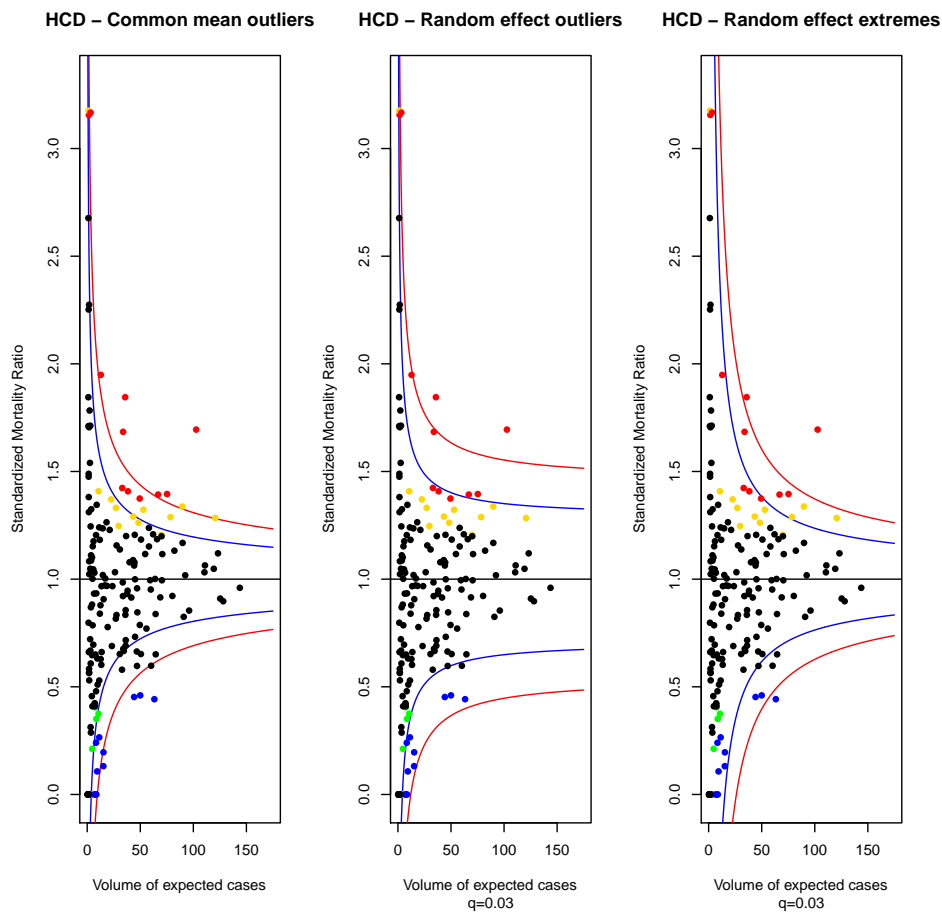


Figura 7.15: Funnel Plot basati sul modello *random effect*, con evidenziate le strutture affette da effetti casuali notevoli ($q = 0.03$).

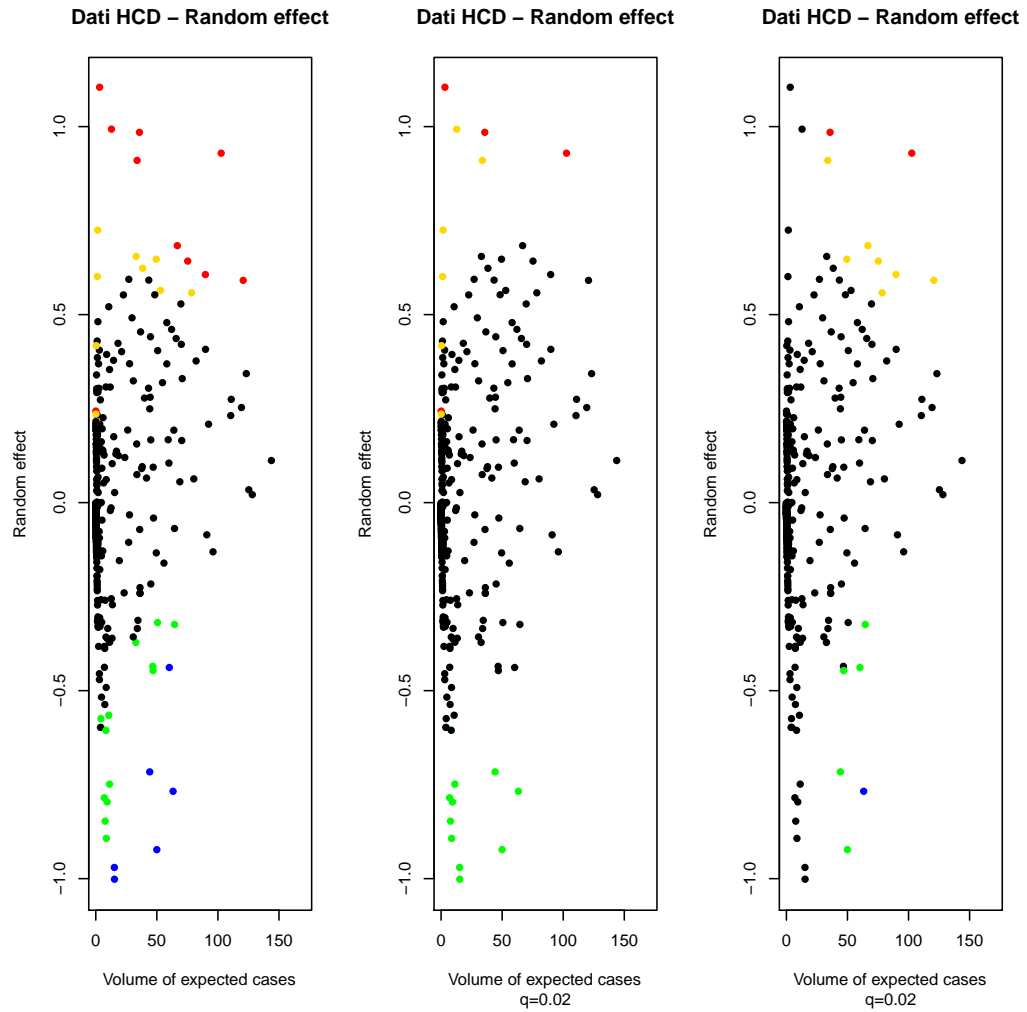


Figura 7.16: Scatter plot con evidenziate le unità indicate dai relativi Funnel Plot (common mean outliers a sinistra, random effect outliers al centro e random effect extremes a destra). Winsorization pari a $q = 0.02$.

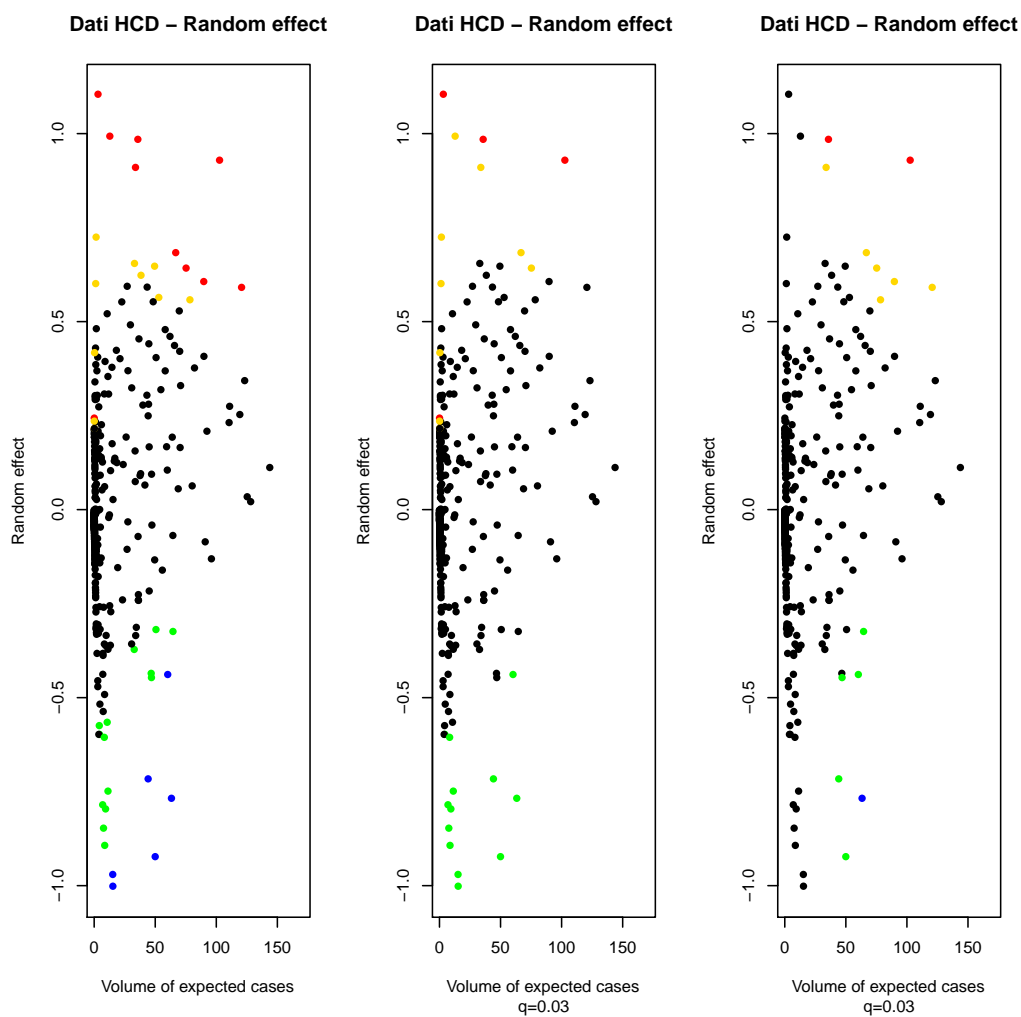


Figura 7.17: Scatter plot con evidenziate le unità indicate dai relativi Funnel Plot (common mean outliers a sinistra, random effect outliers al centro e random effect extremes a destra). Winsorization pari a $q = 0.03$.

strutture evidenziate dai relativi Funnel Plot sono molto sparse, anche di più che nei grafici relativi agli outliers dal modello common mean.

Ritroviamo quindi quanto detto nella sezione 7.3.2, ovvero che i grafici costruiti con la correzione additiva, equivalenti ai grafici per la ricerca di outliers dal modello *random effects*, sono in questo caso quelli più promettenti. Riescono infatti a cogliere abbastanza bene le strutture con effetti sopra il quantile al 97.5% e sotto il quantile al 2.5%. Contemporaneamente alle strutture indicate come interessanti sul Funnel Plot corrisponde un effetto casuale nella fascia superiore o inferiore della distribuzione empirica degli stessi, evidenziando due nuvole abbastanza compatte sul grafico degli effetti casuali stimati. Uniche eccezioni, in entrambe le figure si possono osservare un punto evidenziato in rosso più in basso a sinistra (corrispondente all'ospedale col codice 13, al quale è associato un solo paziente) e altre tre strutture evidenziate in giallo a destra; nella Figura 7.17 si può osservare nella fascia bassa una struttura indicata in verde lievemente più in alto delle altre.

7.5 Confronto con il QQ-plot degli effetti casuali

Da quanto visto nelle precedenti sezioni, il Funnel Plot che ha presentato il comportamento migliore rispetto al raffronto col box-plot degli effetti casuali è stato quello con la correzione additiva (o grafico per la ricerca degli outlier dal modello *random effect*) con livello di winsorization tra $q = 0.02$ e $q = 0.03$. Per concludere sarebbe interessante vedere come si posizionano le strutture segnalate sul QQ-plot degli effetti casuali; teniamo sempre a mente infatti che abbiamo assunto per gli effetti casuali degli ospedali una distribuzione Gaussiana.

Come già ci aspetteremmo, per quanto emerso dagli scatter plot degli effetti casuali mostrati in precedenza, le strutture segnalate dai Funnel Plot corrispondono agli effetti nelle code dei grafici in Figura 7.18 e 7.19. Dalla forma del QQ-plot si può vedere come siano presenti in effetti delle deviazioni dalla normalità. Conducendo un test di normalità otteniamo un p-value molto basso; usando il test di Shapiro abbiamo infatti:

Shapiro-Wilk normality test

```
data: ranef
W = 0.9376, p-value = 3.968e-12
```

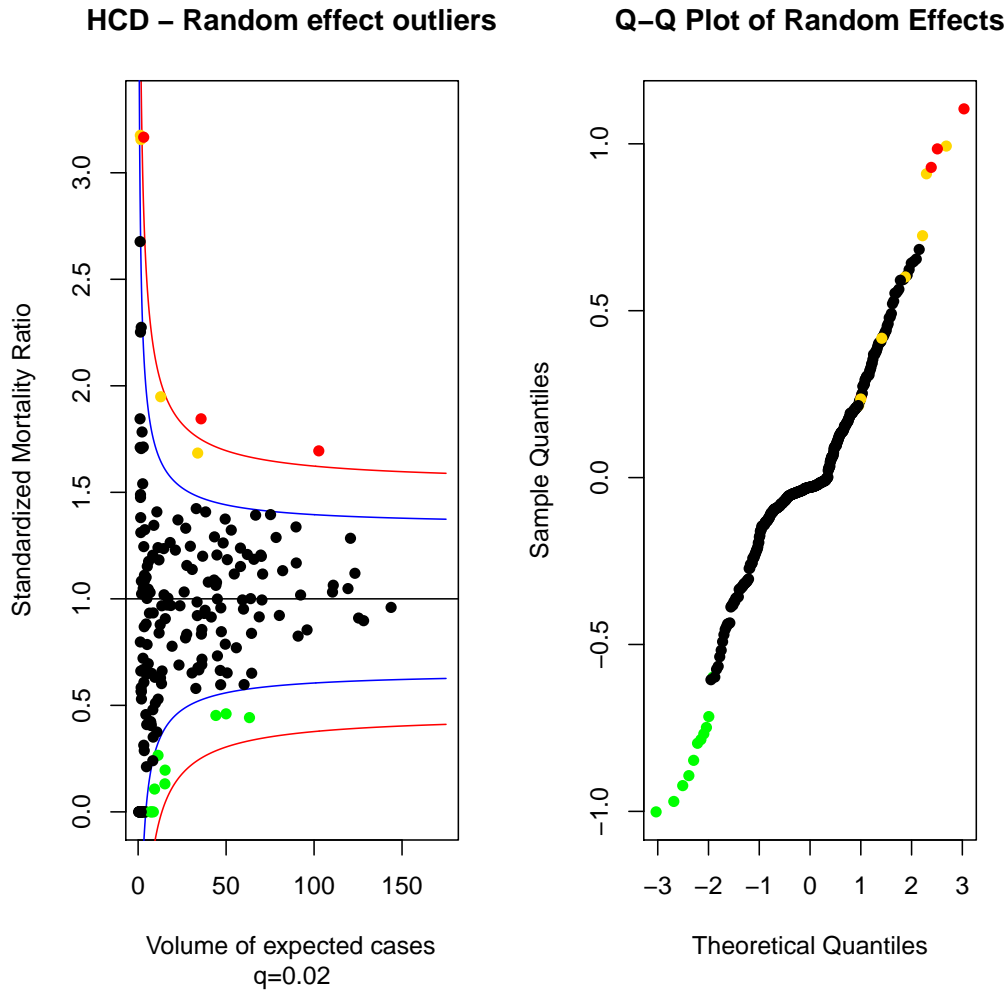


Figura 7.18: Funnel Plot per l'SMR con correzione additiva ($q = 0.02$) a sinistra e QQ-plot degli effetti casuali a destra. Su tutti i grafici sono evidenziati gli ospedali indicati dal Funnel Plot. In particolare in rosso gli ospedali oltre il limite al 99.9% e in blu gli effetti sotto il limite al 0.1%.

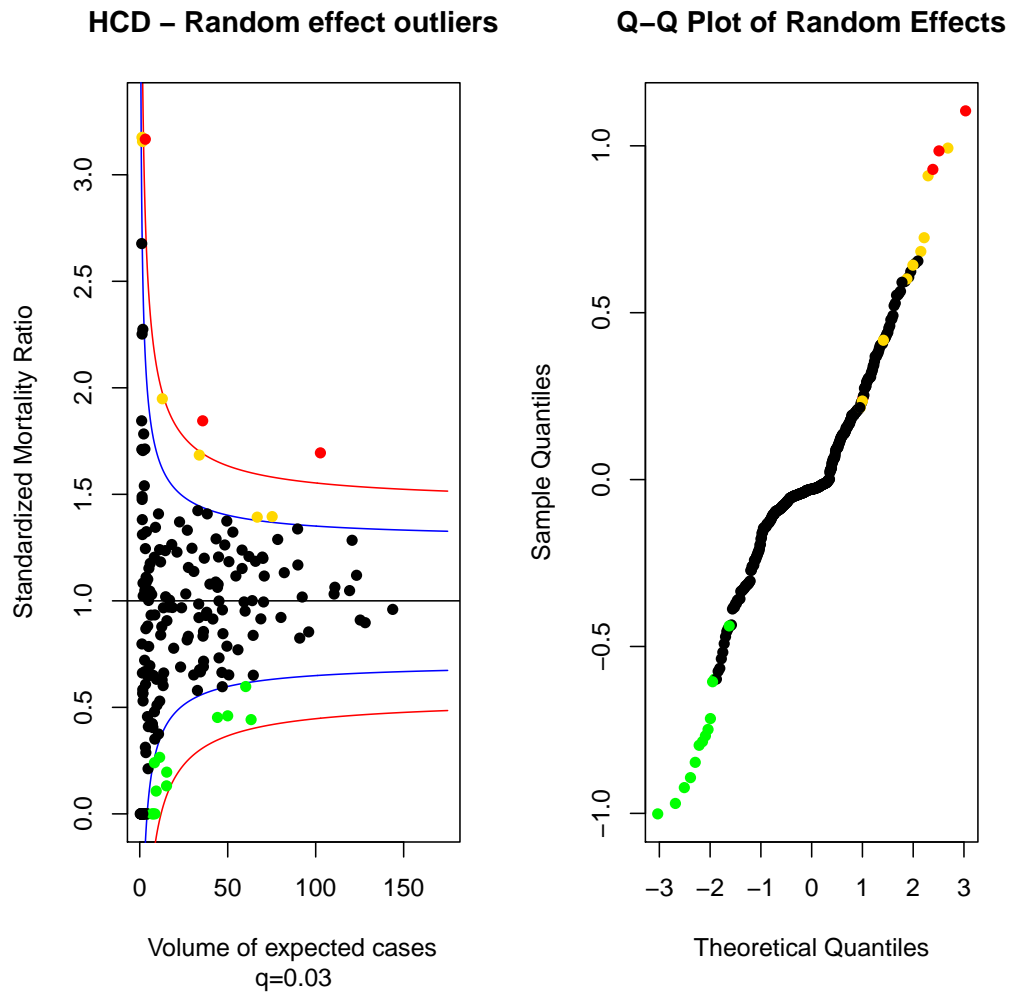


Figura 7.19: Funnel Plot per l'SMR con correzione additiva ($q = 0.03$) a sinistra e QQ-plot degli effetti casuali a destra. Su tutti i grafici sono evidenziati gli ospedali indicati dal Funnel Plot. In particolare in rosso gli ospedali oltre il limite al 99.9% e in blu gli effetti sotto il limite al 0.1%.

che ci porta al rifiuto dell'ipotesi di normalità per gli effetti casuali. Levando le strutture segnalate dal Funnel Plot di Figura 7.19 come outliers la situazione non migliora; con un altro test abbiamo:

Shapiro-Wilk normality test

```
data: ranef[-which(color != "black")]  
W = 0.9545, p-value = 1.509e-09
```

che ci porta ancora al rifiuto dell'ipotesi di normalità.

Concludendo, i Funnel Plot costruiti si sono comportati bene riguardo alla segnalazione delle strutture con effetti più estremi. Per risolvere i problemi di normalità degli effetti non è sufficiente in questo caso eliminare le strutture indicate. È probabilmente necessario procedere per altre vie (trasformazioni dei dati, etc.).

Conclusioni

ABBIAMO presentato in questo lavoro la metodologia di analisi grafica basata sui Funnel Plot. Le loro principali caratteristiche, come rimarcato nei primi capitoli, sono essenzialmente le seguenti:

- Sono semplici da leggere: l'occhio viene automaticamente guidato verso le strutture interessanti;
- Sono semplici da costruire: non è necessario costruire un intervallo di confidenza per ogni struttura in esame, cosa che sarebbe richiesta per costruire un grafico caterpillar. I limiti di confidenza, che non dipendono dai dati e che risultano maggiormente agevoli da costruire, ne prendono il posto;
- Sono notevolmente versatili: si possono applicare a varie situazioni attraverso l'utilizzo del modello adeguato (nel Capitolo 1 ne vengono presentati alcuni dei più utili).

Un tale strumento quindi può trovare certamente utilissime applicazioni nella pratica, allo scopo di monitorare numerosi fenomeni e processi, in particolare una volta che sia stato gestito il problema dell'over-dispersion (Capitoli 2 e 3). I fenomeni di interesse possono essere studiati sia da un punto di vista 'istantaneo' o seguiti nella loro evoluzione temporale attraverso un'opportuna definizione dell'indice di performance.

Nel seguito abbiamo presentato i modelli GLM e GLMM (Capitoli 5 e 6). Ciò che abbiamo potuto apprezzare è il forte legame tra le stime degli effetti casuali provenienti da un modelli di tipo GLMM e i Funnel Plot costruiti basando l'indice di performance e l'indice di previsione sul relativo modello senza effetti casuali (GLM), in particolare i Funnel Plot costruiti con la correzione additiva (o grafici per la ricerca degli outliers descritti in dettaglio nel Capitolo 3).

In questo lavoro ci si è occupati preminentemente delle possibili applicazioni in ambito clinico (Capitolo 4 e 7), applicando le metodologie descritte nei primi tre capitoli al data set amministrativo di Regione Lombardia per

il monitoraggio delle morti intra-ospedaliere a seguito di problemi cardiovascolari negli ospedali della regione. Le autorità preposte al monitoraggio del comportamento di queste strutture potrebbero utilizzare tali metodologie come uno strumento aggiuntivo per identificare strutture interessanti in positivo e in negativo, seguendone i risultati per lunghi periodi di tempo, magari analizzando un'opportuna finestra temporale che slitti in avanti di anno in anno.

Va da sè che questi metodi grafici potrebbero trovare tranquillamente applicazione in numerosissimi altri campi medici, ad esempio per seguire le performance di cliniche specializzate in settori specifici (non necessariamente per disturbi di tipo cardiaco) o per monitorare le grandi strutture ospedaliere. Allo stesso modo si potrà approfondire l'utilizzo di questo paradigma anche all'esterno di questi campi, ad esempio per giudicare strutture scolastiche valutando le bocciature durante un anno scolastico, la quantità di studenti ammessi agli esami di maturità e molto altro ancora. Insomma, in conclusione, appare evidente la grande versatilità di queste procedure grafiche.

Ringraziamenti

I DATI relativi ai ricoveri di pazienti affetti da patologie cardiovascolari utilizzati in questo lavoro sono stati resi disponibili all'interno del progetto di 'Ricerca Finalizzata Ministero della Salute' dell'anno 2009 **Utilization of regional health service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure**, e provengono dalla *Banca dati Assistito*, il database amministrativo di Regione Lombardia. Si rivolge un sincero ringraziamento per aver reso possibile l'utilizzo di queste informazioni.

Si ringrazia la Dottoressa Francesca Ieva per la sua disponibilità e per aver messo a disposizione le sue conoscenze e il suo studio pregresso del data set relativo ai ricoveri per patologie cardiovascolari. Il suo contributo a questo lavoro è stato di grandissima importanza.

Un sincero ringraziamento va alla Professoressa Anna Maria Paganoni, per il tempo dedicatomi nella stesura di questa Tesi e per la professionalità con cui mi ha seguito e indirizzato durante tutto il periodo di lavoro.

Infine un grazie ai miei familiari: la mamma, il papà e Ilaria.

Ah, e da ultimo, ma non per ultimo, un grande, grandissimo grazie anche a Silvia ♡.

Bibliografia

- [1] Agresti A., *Categorical Data Analysis* (2nd Edition), John Wiley & Sons, 2002.
- [2] Burgess J. F., Christiansen C. L., Michalak S. E., Morris C. N., *Medical Profiling: Improving Standards and Risk Adjustment Using Hierarchical Models*, Journal of Health Economics, 19, 291-309, 2000.
- [3] Christiansen C. L., Morris C. N., *Hierarchical Poisson Regression Modeling*, Journal of the American Statistical Association, 92, 618-632, 1997.
- [4] Darlow B. A., Hutchinson J. L., Simpson J. M., Henderson-Smart D. J., Donoghue D. A., Evans N. J., *Variation in Rates of Severe Retinopathy of Prematurity Among Neonatal Intensive Care Units in the Australian and New Zealand Neonatal Network*, British Journal of Ophthalmology, 89, 1592-1596, 2005.
- [5] Goldstein H., *Multilevel Statistical Models*, (3rd Edition), London: Edward Arnold, 2003.
- [6] Jones H. E., Spiegelhalter D. J., *The Identification of 'Unusual' Health-Care Providers From a Hierarchical Model*, The American Statistical Association, 2011, vol 65(3), 154-163.
- [7] McCullagh P., Nelder J. A., *Generalized Linear Models* (2nd Edition), Chapman & Hall, London, 1989.
- [8] Montgomery D. C., Runger G. C., Hubele N. F., *Statistica per ingegneria*, Egea, 2004.
- [9] Morris C. N., Christiansen C. L., *Hierarchical Models for Ranking and for Identifying Extremes, With Applications*, Oxford University Press, 277-296, 1996.
- [10] Ross S. M., *Calcolo delle probabilità*, Apogeo, 2007.

- [11] Simpson J. M., Evans N. J., Gibbert R. W., Heuchan A. M., Henderson-Smart D. J., *Analysing Differences in Clinical Outcomes Between Hospitals*, Quality and Safety in Health Care, 12, 257-262, 2003.
- [12] Smits J. M. A., Meester J. D., Deng M. C., Scheld H. H., Hummel M., Schoendube F., Haverich A., Vanhaecke J., Van Houwelingen H. C., *Mortality Rates After Heart Transplantation: How to Compare Center-Specific Outcomes Data?*, Transplantation, 75, 90-96, 2003.
- [13] Spiegelhalter D. J., *Funnel plots for comparing institutional performance*, Statistics in Medicine 2005; 24;1185-1202.
- [14] Spiegelhalter D. J., *Funnel plots for institutional comparisons (letter)*, Quality and Safety in Health Care 2002; 11; 390-391.
- [15] Spiegelhalter D. J., *Handling Over-Dispersion of Performance Indicators*, Quality and Safety in Health Care 2005; 14; 347-351.
- [16] Racz M. J., Sedransk J., *Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes*, Journal of the American Statistical Association, 105, 48-58, 2010.
- [17] *Coronary Artery Bypass Surgery in New York State, 1996-1998*, New York State Department of Health, 2001.
- [18] *Coronary Artery Bypass Surgery in New York State, 1997-1999*, New York State Department of Health, 2002.