UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA, ECONOMIA E FINANZA

TESI DI LAUREA TRIENNALE

UNA APPLICAZIONE DEL MODELLO DI POISSON CON INFLAZIONE DI ZERI

Relatore: Ch.ma Prof.ssa Laura Ventura

Laureanda: Sara Baldo

Matr. N. 527073- SEF

ANNO ACCADEMICO 2006/2007

Indice

pagina

Introduzione

Il modello di regressione di Poisson viene ampiamente utilizzato per l'analisi di dati di tipo conteggio. In alcune applicazioni, tuttavia, questo tipo di dati possono mostrare una sostanziale sovradispersione, che si manifesta quando i dati hanno una maggiore variabilità rispetto a quanto ammesso dal modello. In particolare, l'uguaglianza di media e varianza per i dati analizzati, sotto le assunzioni di un modello di regressione di Poisson, viene violata. Varie ragioni, quali ad esempio una forte eterogeneità, covariate mancanti o correlazioni tra misurazioni, possono condurre alla sovradispersione (si vedano Cameron e Trivedi, 1998, o Winkelmann, 2000). Di conseguenza, la letteratura statistica ha proposto varie estensioni del modello di regressione di Poisson, adatte a trattare situazioni di concentrazione eccessiva di zeri nei dati. In questo contesto, si desidera approfondire il modello di Poisson con inflazione di zeri (*Zero-Inflated Poisson regression*).

In generale, per una variabile di tipo conteggio è conveniente usare la regressione di Poisson. Per $i = 1,\ldots, n$, si assume che $Y_i \sim \text{Poisson}(\lambda)$, dove il valore atteso e la varianza sono pari a λ, ossia $E(Y_i) = \text{Var}(Y_i) = \lambda$. Qualora i dati presentino molti valori pari a zero, può essere inopportuno adattare il modello di Poisson in quanto, anche tramite una semplice analisi esplorativa, si può confutare l'uguaglianza di media e varianza. Infatti, indicata con λ la media della distribuzione di Poisson, un campione di numerosità n elevata dovrebbe avere circa $ne^{-\lambda}$ unità statistiche che assumono valore zero. Qualche volta, tuttavia, ci sono molte più unità statistiche che assumono valore zero rispetto a quelle che sarebbero state predette dal modello poissoniano che si limita a modellare, dunque a prevedere, valori positivi ma non è predisposto a modellare dati che presentano un'eccessiva presenza di zeri.

- ***Esempio 1***

  Lambert (1992) suggerisce che, avvalendosi di un processo di manifattura, quando un processo affidabile è sotto controllo, il numero dei difetti dell'oggetto preso in considerazione dovrebbe seguire la distribuzione di Poisson. Indicata con λ la media della distribuzione di Poisson, un campione di numerosità n elevata dovrebbe avere circa $ne^{-\lambda}$ unità statistiche con nessun difetto. Qualche volta, tuttavia, ci sono molte più unità statistiche senza difetti rispetto a quelle che sarebbero state predette dal modello, basandosi sul numero di difetti delle unità imperfette. Un'interpretazione è che cambiamenti lievi e inosservati del processo produttivo causano un movimento casuale oscillante tra un perfetto stato, nel quale i difetti sono estremamente rari, e uno stato in cui i difetti sono possibili ma non inevitabili. Il perfetto stato di transizione, cioè il momento in cui ci sono *items* che sono eccezionalmente senza difetti, contribuisce ad aumentare il numero di zeri nei dati.

Il modello di Poisson con inflazione di zeri (ZIP) nasce proprio per trattare dati di tipo conteggio che si infittiscono in particolar modo attorno allo zero. Questi modelli senza covariate sono stati discussi da vari autori (ad esempio, si vedano Cohen, 1963; Johnson e Kots, 1969), ma in questo contesto considereremo il caso in cui, detta p la probabilità di avere una variabile risposta pari a zero e λ il valore atteso della distribuzione di Poisson, entrambi i parametri p e λ possono dipendere dalle covariate. In particolare, le quantità $\log(\lambda)$ e $\text{logit}(p) = \log\frac{p}{1-p}$ sono assunti come funzioni lineari di alcune covariate. Le stesse (o diverse) covariate possono influire su p e λ, e i parametri p e λ potrebbero o non potrebbero essere funzionalmente collegati. Quando p è una funzione decrescente di λ, la probabilità del perfetto stato e la media dell'imperfetto stato aumentano o diminuiscono insieme.

Si ricorda anche Heilbron (1989) che nello stesso periodo ha proposto un modello di regressione simile, chiamato *zero-altered Poisson*. Sebbene i modelli siano stati sviluppati indipendentemente, l'acronimo ZIP è una modifica adattata dell'acronimo di Heilbron ZAP, che sta per *zero-altered Poisson*. Anche altri autori hanno

precedentemente considerato la fusione di una distribuzione che degenera intorno allo 0, con una distribuzione a parte, che può essere la Poisson o la binomiale negativa.
Lo schema della tesi è il seguente. Nel Capitolo 1 viene presentato il modello ZIP e viene accennata l'inferenza basata sulla verosimiglianza. Nel Capitolo 2 sono presentate le procedure e i comandi del pacchetto statistico `pscl` del *freesoftware* `R` che permettono di adattare il modello ZIP ad un insieme di dati. Infine, nel Capitolo 3 si discute un'applicazione del modello ZIP ad un insieme di dati di natura epidemiologica.

CAPITOLO 1

Il Modello di Poisson con Inflazione di Zeri

1.1 Introduzione

In questo capitolo viene introdotto il modello di regressione poissoniano con inflazione di zeri, assieme alle sue caratteristiche fondamentali, dalla presentazione generale del modello, all'inferenza sulla verosimiglianza, e per concludere, agli intervalli di confidenza per i parametri del modello. L'articolo più completo nel quale è possibile reperire informazioni riguardo questo modello è il lavoro di Lambert (1992). Inoltre, una trattazione sul modello può essere trovata anche in Cameron e Trivedi (1998).

1.2 Il Modello

Il modello di regressione di Poisson con inflazione di zeri, è stato proposto da Lambert (1992) per trattare situazioni con eccesso di zeri, ossia quando la maggior parte dei dati si concentra intorno allo zero. L'idea di base è che una probabilità binomiale governa l'esito binario di una realizzazione dell'evento nello zero o tra valori positivi. In altre parole, il processo che genera i dati ha due stati: il primo è uno stato in cui sono osservati solo valori praticamente nulli e il secondo si riferisce a una distribuzione di Poisson o a una binomiale in cui sono osservati valori diversi da zero e pochi zeri.

Questo modello, dunque, contribuisce a risolvere il problema della eccessiva concentrazione dei dati intorno allo zero.
Nel modello di regressione ZIP, le variabili risposta $Y = (Y_1, \ldots, Y_n)^T$ sono indipendenti e tali che

$Y_i \sim 0$ con probabilità p_i
$Y_i \sim \text{Poisson}(\lambda_i)$ con probabilità $(1 - p_i)$, $i = 1, \ldots, n$,

così che

$Y_i = 0$ con probabilità $p_i + (1 - p_i)e^{-\lambda i}$
$Y_i = k$ con probabilità $(1 - p_i)e^{-\lambda i}\lambda_i^k/k!$, con $k = 1,2, \ldots$.

Inoltre, i parametri $\lambda = (\lambda_1, \ldots, \lambda_n)^T$ e $p = (p_1, \ldots, p_n)^T$ possono dipendere da covariate e sono modellati per mezzo dei legami canonici per il modello di Poisson e per il modello binomiale. Pertanto, i parametri soddisfano le seguenti uguaglianze:

$$\log(\lambda) = B\beta$$

e (1)

$$\text{logit}(p) = \log\frac{p}{1-p} = G\gamma,$$

con B e G matrici di regressione non necessariamente uguali, λ media della distribuzione di Poisson e p probabilità che la variabile risposta dia esito pari a zero. Dato che i parametri di interesse per interpretare il modello ZIP sono λ e p, le uguaglianze in (1) possono essere invertite in modo tale da ottenere l'espressione esplicita per i parametri, ossia

$$\lambda = \exp(B\beta)$$

e (2)

$$p = \frac{\exp(\mathrm{G}\gamma)}{1+\exp(\mathrm{G}\gamma)}.$$

La media e la varianza di Y_i, $i = 1, \ldots, n$ sono, rispettivamente,

$$\mathrm{E}(Y_i) = (1 - p_i)\,\lambda_i \tag{3}$$

e

$$\mathrm{Var}(Y_i) = (1 - p_i)\,\lambda_i\,(1 + p_i\,\lambda_i),\ i = 1, \ldots, n. \tag{4}$$

Sebbene questo modello consista in due parti distinte, le due componenti del modello devono essere adattate ai dati contemporaneamente. Heilbron (1994) propone un modello alternativo diviso in due parti, che permette alle due componenti del modello di essere stimate separatamente, ma questo tipo di stima può essere utile quando si deve scegliere tra una moltitudine di distribuzioni oltre che alla Poisson.

Nel modello di Lambert (1992), le matrici B e G contengono potenzialmente differenti insiemi di covariate, che riguardano sia la probabilità dello stato in cui vi sono gli zeri (che corrisponde alla probabilità di successo) sia la media della distribuzione di Poisson nello stato in cui non vi sono gli zeri (che corrisponde alla probabilità di insuccesso). Perciò γ viene interpretato come l'effetto del livello dei fattori o delle covariate sulla probabilità di successo e l'interpretazione di β è legata all'effetto sulla media del numero di insuccessi. Una semplice conseguenza del modello ZIP è che, poiché marginalmente $\mathrm{E}(Y_i) = (1 - p_i)\,\lambda_i$, il numero medio dei valori assunti dalla variabile risposta Y_i dipende da γ attraverso p_i e da β attraverso λ_i. Perciò, per una covariata che capita in entrambe le matrici G e B, i test di verifica di ipotesi dicono che sulla base di ciò, i valori γ e β delle covariate sono significativi quando i test sono condotti separatamente o contemporaneamente (si veda Hall, 2000).

Le covariate che incidono sulla media della Poisson possono essere le stesse (oppure no) delle covariate che influiscono sulla probabilità p. Quando le covariate coincidono e λ e p non sono funzionalmente relazionate, si ha che B = G e la regressione ZIP richiede un numero doppio di parametri della regressione di Poisson. Al contrario, quando la

probabilità p non dipende dalle covariate, G è un vettore di 1, e la regressione ZIP richiede solo un parametro in più della regressione di Poisson.

Se le stesse covariate influiscono su p e λ, è naturale ridurre il numero di parametri pensando a p come una funzione di λ. Assumendo che la funzione sia nota a meno di una costante, si semplifica di quasi la metà il numero dei parametri di cui necessita la regressione ZIP e questo può accelerare notevolmente i calcoli (si veda Chambers, 1977, p. 144). In molte applicazioni, tuttavia, ci può essere un'informazione un po' più precisa su come p è legata a λ. In questo caso, una naturale parametrizzazione è

$$\log(\lambda) = \mathrm{B}\beta$$

e (5)

$$\mathrm{logit}(p) = -\tau\mathrm{B}\beta,$$

con τ $(-\infty < \tau < +\infty)$, che rappresenta un parametro di forma ignoto, che implica che $p_i = (1 + \lambda_i^{\tau})^{-1}$, per $i = 1,\ldots, n$.

Nel linguaggio dei modelli lineari generalizzati, log(λ) e logit(p) sono i legami canonici o le trasformazioni che rendono lineari la media della Poisson e la probabilità di successo della bernoulliana. Se il termine Bβ è pensato come un vincolo, i legami canonici per p e λ sono entrambi proporzionali a Bβ. I modelli ZIP con il legame logit per p, il legame logaritmo per λ e con il parametro di forma τ (4), saranno indicati con la sigla ZIP(τ).

Il legame logit per p è simmetrico intorno allo 0.5. Due diffusi legami asimmetrici sono il legame log-log definito come $\log(-\log(p)) = \tau\mathrm{B}\beta$ o, equivalentemente, $p_i = \exp(-\lambda_i^{\tau})$ e il legame complementare log-log, definito da $\log(-\log(1 - p)) = -\tau\mathrm{B}\beta$, oppure $p_i = 1 - \exp(-\lambda_i^{-\tau})$. Heilbron (1989) ha usato inoltre un legame definito come $p_i = \exp(-\tau\lambda_i)$, oppure $\log(-\log(p)) = \mathrm{B}\beta + \log(\tau)$. Altri legami logit lineari e i legami log-log potrebbero essere definiti da $\mathrm{logit}(p) = \log(\alpha) - \tau\mathrm{B}\beta$ e $\log(-\log(p)) = \log(\alpha) + \tau\mathrm{B}\beta$, rispettivamente, con $\alpha > 0$.

Con uno qualsiasi di questi legami, quando $\tau > 0$, lo zero diventa meno probabile quando la media λ_i si incrementa; invece quando $\tau \to \infty$ lo zero diventa impossibile se λ_i è fissato. Quando $\tau \to 0$, sotto il legame additivo log-log e quando $\tau \to -\infty$ sotto

l'altro legame, lo zero diventa certo. Un τ negativo non è permesso usarlo con il legame additivo log-log, ma per gli altri legami con $\tau < 0$, la media della Poisson aumenta tanto quanto l'eccesso di zeri diventa più probabile.

1.3 Stima di Massima Verosimiglianza

Il numero di parametri che possono essere stimati in un modello di regressione ZIP dipende dalla numerosità campionaria. Se ci sono solo pochi dati positivi e λ e p non sono legati funzionalmente, allora solo semplici modelli dovrebbero essere considerati per la componente del modello di Poisson, ossia λ. I dati sono adeguati per la stima dei parametri di un modello ZIP o ZIP(τ) se l'informazione osservata è una matrice non singolare (si veda tale matrice nell'Appendice in Lambert, 1992).

1.3.1 λ e p Indipendenti

Quando λ e p non sono funzionalmente relazionate, la log-verosimiglianza per la regressione ZIP con la parametrizzazione standard (1) è

$$\begin{aligned} L(\gamma, \beta; y) = & \sum_{y_i=0} \log(e^{G_i\gamma} + \exp(-e^{B_i\beta})) \\ & + \sum_{y_i>0} (y_i B_i\beta - e^{B_i\beta}) \\ & - \sum_{i=1}^{n} \log(1 + e^{G_i\gamma}) \\ & - \sum_{y_i>0} \log(y_i!), \end{aligned} \qquad (6)$$

dove G_i e B_i sono le righe i-esime delle matrici G e B. La somma delle funzioni esponenziali nel primo termine complica la massimizzazione della funzione di log-verosimiglianza. Ma se si suppone di conoscere quali zeri provengono dal perfetto stato

e quali provengono dalla Poisson, ossia se si suppone di poter disporre di una variabile Z che assume valori 0 e 1, tale che $Z_i = 1$ quando Y_i è nello stato perfetto e $Z_i = 0$ quando Y_i viene dallo stato imperfetto, ovvero dalla Poisson, allora la log-verosimiglianza completa per (γ, β), denotata con $L_C(\gamma, \beta; y, z)$, che dipende da y e z, diventa

$$L_C(\gamma, \beta; y, z) = \sum_{i=1}^{n} \log(f(z_i|\gamma)) + \sum_{i=1}^{n} \log(f(y_i| z_i, \beta))$$

$$= \sum_{i=1}^{n} (z_i G_i\gamma - \log(1 + e^{G_i\gamma})) + \sum_{i=1}^{n} (1 - z_i)(y_i B_i\beta - e^{B_i\beta}) - \sum_{i=1}^{n} (1-z_i) \log(y_i!)$$

$$= L_C(\gamma; y, z) + L_C(\beta; y, z) - \sum_{i=1}^{n} (1 - z_i) \log(y_i !).$$

Questa log-verosimiglianza è facile da massimizzare, in quanto $L_C(\beta; y, z)$ e $L_C(\gamma; y, z)$ possono essere massimizzate separatamente.

Con l'algoritmo EM (per un approfondimento si veda Xiao-Li Meng, 2004), che richiede tre fasi, la log-verosimiglianza (6) viene massimizzata iterativamente alternando la stima di Z_i con il suo valore atteso condizionato alle stime correnti di (γ, β) (E step) e, successivamente, considerando costanti i valori di Z_i ottenuti nel passo precedente, la massimizzazione di $L_C(\gamma, \beta; y, z)$ (M-step). Per lo svolgimento dettagliato dell'algoritmo EM si veda Lambert (1992). Le stime che si ottengono nell'ultima iterazione sono le stime di massima verosimiglianza $(\hat{\gamma}, \hat{\beta})$ per (γ, β).

Ci sono altri algoritmi, come l'algoritmo di Newton-Raphson, per massimizzare la log-verosimiglianza (6). Quando converge, l'algoritmo di Newton-Raphson è di solito più veloce dell'algoritmo EM. L'algoritmo EM, tuttavia, può essere più facile da programmare.

1.3.2 Il Parametro *p* come Funzione di λ

La log-verosimiglianza per il modello ZIP(τ) con la parametrizzazione standard (2) è, a meno di una costante,

$$L(\beta, \tau, y) = \sum_{y_i=0} \log(e^{-\tau B_i\beta} + \exp(-e^{B_i\beta})) + \sum_{y_i>0} (y_i B_i\beta - e^{B_i\beta}) - \sum_{i=1}^{n} \log(1 + e^{-\tau B_i\beta}). \quad (7)$$

L'algoritmo EM non è utile in questo caso, perché i parametri β e τ non possono essere stimati con semplicità anche se Z è noto. L'algoritmo di Newton-Raphson, in questo caso, può essere preferibile. Una stima iniziale nell'algoritmo può essere $\beta^{(0)} = \hat{\beta}_u$ per β e $\tau^{(0)} = -$ mediana $(\hat{\gamma}_u, \hat{\beta}_u)$ per τ, dove $(\hat{\gamma}_u, \hat{\beta}_u)$ sono le stime di massima verosimiglianza del modello ZIP.

1.4 Standard Error e Intervalli di Confidenza

Per campioni di dimensione elevata, gli stimatori di massima verosimiglianza dei parametri $(\hat{\gamma}, \hat{\beta})$ per il modello ZIP e dei parametri $(\hat{\beta}, \hat{\tau})$ per il modello ZIP(τ) si distribuiscono approssimativamente come normali con medie (γ, β) e (β, τ), rispettivamente, e varianze uguali alle inverse delle rispettive matrici dell'informazione osservata di Fisher. Pertanto, per n sufficientemente elevato, gli stimatori di massima verosimiglianza e funzioni regolari di questi, come ad esempio la probabilità p di ottenere degli zeri o la media λ della distribuzione di Poisson, sono approssimativamente non distorti.

Gli intervalli di confidenza alla Wald, sono facili da costruire, ma essi assumono che la log-verosimiglianza sia approssimativamente quadratica attorno al massimo della verosimiglianza. In alternativa, vari studi di simulazione (si veda Lambert, 1992), suggeriscono che gli intervalli di confidenza basati sul log-rapporto di verosimiglianza, più difficili da calcolare, sono tuttavia più affidabili.

Un intervallo di confidenza di livello approssimato $(1 - \alpha)$ per un qualunque parametro di regressione β_j nella regressione ZIP, per esempio, si trova utilizzando la statistica log-rapporto di verosimiglianza profilo, data da

$$W_p(\beta_j) = 2(l_p(\hat{\beta}_j) - l_p(\beta_j)), \quad (8)$$

dove $l_p(\beta_j)$ è la funzione di log-verosimiglianza profilo per il parametro β_j, che si ottiene sostituendo nella funzione di log-verosimiglianza (6), i parametri di disturbo (γ e tutti gli altri elementi di β) con la loro stima di massima verosimiglianza per β_j fissato, e $l_p(\hat{\beta}_j)$ è la stima della funzione di log-verosimiglianza, che si ottiene sostituendo nella funzione di log-verosimiglianza (6), i valori stimati dei parametri β e γ. Naturalmente, bisogna confrontare la suddetta quantità (8) con $\chi^2_{1,(1-\alpha)}$, che rappresenta il quantile di livello $(1 - \alpha)$ di una distribuzione χ^2 con 1 grado di libertà. L'intervallo di confidenza per β_j è individuato dalla seguente disuguaglianza:

$$W_p(\beta_j) < \chi^2_{1,(1-\alpha)}. \qquad (9)$$

Analogo è, ovviamente, il processo per individuare un intervallo di confidenza per una componente γ_j dell'altro parametro di regressione γ, nel modello ZIP.
Inoltre, per n sufficientemente elevato, risulta affidabile stimare intervalli di confidenza alla Wald. Prendendo in considerazione un qualsiasi coefficiente, denotato con β_j, l'intervallo di confidenza risulta essere centrato sul valore $\hat{\beta}_j$, stima di massima verosimiglianza del parametro, ossia

$$\hat{\beta}_j \pm z_{\left(1-\frac{\alpha}{2}\right)} \; j_p\left(\hat{\beta}_j\right)^{-\frac{1}{2}} \quad , \qquad (10)$$

dove $z_{(1 - \alpha/2)}$ è il quantile di livello $(1 - \alpha/2)$ della funzione di una normale-standard, $j_p\left(\hat{\beta}_j\right)^{-\frac{1}{2}}$ è l'inversa dell'informazione profilo osservata calcolata a partire dalla $l_p(\beta_j)$.

1.5 Test e Verifiche di Ipotesi

Basandosi sulla teoria della verosimiglianza, è noto che per la verifica d'ipotesi per testare se un parametro generico β_j può assumere un qualsiasi valore β_0, ossia H_0: $\beta_j = \beta_0$, si utilizza la statistica test

$$T = \frac{\hat{\beta}_j - \beta_0}{s.e.\left(\hat{\beta}_j\right)} = j_p\left(\hat{\beta}_j\right)^{-\frac{1}{2}}\left(\hat{\beta}_j - \beta_0\right), \qquad (11)$$

con $\hat{\beta}_j$ stima di massima verosimiglianza del parametro β_j, β_0 il valore per il quale si desidera saggiare l'ipotesi H_0, e $s.e.(\hat{\beta}_j)$ lo standard error, che equivale alla radice quadrata del reciproco dell'informazione osservata (si veda Pace e Salvan, 2001) calcolata a partire dalla funzione $l_p(\beta_j)$. Si indichi con t^{oss} la statistica test calcolata sulla base dei dati. La seguente tabella riporta le procedure per valutare come accettare H_0 e calcolare il relativo *p-value*:

Alternativa	*Sono significativi contro H_0*	*Regione di rifiuto del test con livello α approssimato*	*α^{oss} approssimato*
H_1^{dx}: $\beta > \beta_0$	Valori grandi di t^{oss}	$t^{oss} > z_{(1-\alpha)}$	$1 - \Phi(t^{oss})$
H_1^{sx}: $\beta < \beta_0$	Valori piccoli di t^{oss}	$t^{oss} < z_{\alpha}$	$\Phi(t^{oss})$
H_1: $\beta \neq \beta_0$	Valori sia grandi sia piccoli di t^{oss}	$\lvert t^{oss}\rvert > z_{(1-\alpha/2)}$	$2\ min\ (\Phi(t^{oss}); 1 - \Phi(t^{oss}))=2(1 - \Phi(\lvert t^{oss}\rvert)$

Tabella 1

In alternativa al test T, si può utilizzare il test $W_p(\beta_j)$ la cui formula (8) è data nel Paragrafo 1.4.

1.6 Conclusioni

In questo capitolo si è introdotto il modello ZIP e si è presentato un breve *excursus* sulle stime dei parametri e sugli intervalli di confidenza per β e γ.

Per riassumere, la regressione ZIP è un modo pratico per modellare i dati di tipo conteggio, che presentano tanti zeri. Il modello è piuttosto facile da programmare e da interpretare. Le stime dei coefficienti, gli standard error basati sull'informazione osservata, il rapporto di log-verosimiglianza per costruire gli intervalli di confidenza, le proprietà delle stime dei parametri (γ, β) e i valori che ci si aspetta nelle previsioni per la variabile risposta sono affidabili. Inoltre, gli stimatori di massima verosimiglianza del

modello ZIP e ZIP(τ) sono asintoticamente normali e due volte la differenza della log-verosimiglianza sotto ipotesi annidate, si distribuisce asintoticamente come un χ^2. Le simulazioni presenti in letteratura suggeriscono che gli intervalli di confidenza basati sul log-rapporto di verosimiglianza sono migliori degli intervalli di confidenza alla Wald; ciò nonostante, sembra che gli intervalli di confidenza alla Wald e i test di verifica di ipotesi funzionino correttamente per β ma sono poco soddisfacenti per γ per valori di $n \leq 100$.

La regressione ZIP si adatta ai dati che si infittiscono attorno allo zero e li mescola con una distribuzione di Poisson. Questo perché, utilizzando il modello ZIP basato sulla distribuzione di Poisson, rende i calcoli più semplici. Infatti, mescolando gli zeri con distribuzioni discrete di famiglie non esponenziali, come la binomiale con una forma ignota, tuttavia, può complicare notevolmente i calcoli (si veda Lambert, 1992).

Alcune applicazioni del modello ZIP partono dal 1992. Si hanno con Lambert (1992) i primi studi su dati reali, riguardanti un processo di manifattura. Tale modello è stato punto di riferimento successivamente per i lavori di Hall (2000) e di van Iersel *et al*. (2000), in cui si accenna allo studio della quantità necessaria di pesticida contro un particolare tipo di insetti.

Nel prossimo capitolo saranno presentati alcuni comandi di R utili quando si devono studiare dati che possiedono i requisiti necessari per l'applicazione del modello ZIP.

CAPITOLO 2

Adattamento del Modello ZIP in R

2.1 Introduzione

Un pacchetto statistico di R da utilizzare per l'adattamento di un modello ZIP, semplice da interpretare, e che si avvale del metodo di stima della massima verosimiglianza (*cfr*. Paragrafo 1.3), è il pacchetto `pscl`, utile per stimare i coefficienti presenti nel modello. In questo capitolo sarà presentata la sintassi di alcuni comandi di R, la cui libreria `pscl`, necessaria per avere a disposizione i comandi per lo studio di tale modello, è scaricabile dalla rete gratuitamente dal sito ufficiale di R, http://www.r-project.org; un *link* utile è http://cran.r-project.org. La sintassi del pacchetto `pscl` relativa al modello ZIP è stata realizzata da Simon Jackman (jackman@stanford.edu). Per usufruire del suddetto pacchetto statistico bisogna caricare la relativa `library`, che raccoglie le funzioni utilizzate per il modello ZIP e si prosegue con i comandi che verranno descritti nei paragrafi seguenti. Ogni paragrafo sarà corredato di un esempio strettamente connesso con quanto spiegato in ciascuno di essi.

2.2 Lettura dei Dati e Sintassi del Pacchetto Statistico `pscl`

Aperta una sessione di R, per usare il pacchetto `pscl` bisogna installare questo insieme di comandi digitando `library(pscl)` ogni volta che si inizia una sessione di lavoro.
In generale, il comando completo per adattare un modello ZIP è il seguente:

```
zeroinfl(count = y ~ ., x = ~1, z = ~1,
         data = list(),
         link = "logit", dist = "poisson",
         method = "BFGS", trace = FALSE, maxit = 50000,
         na.action = na.omit)
```

dove

- ‘`count`’ si riferisce al vettore `y` della variabile risposta di tipo conteggio che rappresenta la parte a sinistra della `formula`. In questo tipo di modello, come accennato nel capitolo precedente, devono eccedere gli zeri.

Invece,

- ‘`x`’ rappresenta la parte destra della `formula` che contiene le covariate per la parte del modello di Poisson di tipo conteggio (secondo la parametrizzazioni precedentemente scritte, `x` indica gli elemento della matrice B; *cfr*. Paragrafo 1.2);
- ‘`z`’ rappresenta la parte destra della `formula` che contiene le covariate per la parte binomiale inflazionata di zeri del modello (analogamente, `z` indica gli elementi della matrice G; *cfr* Paragrafo 1.2);
- ‘`data`‘ contiene il nome del *dataset* (opzionale);
- ‘`link`’ indica il *link* necessario per modellare la probabilità che la variabile assuma valore zero; si riferisce alla parte del modello con inflazione di zeri (la scelta è tra il *link* `logit`, di *default*, e `probit`);
- ‘`dist`’ indica il modello per dati di tipo conteggio: “`poisson`” per default o “`negbin`” (binomiale negativa);
- ‘`method`’ metodo per massimizzare la funzione di log-verosimiglianza (solo “`BFGS`” e “`Nelder-Mead`” sono supportati);
- ‘`trace`’: se `TRUE` il display mostra il processo di massimizzazione della funzione di verosimiglianza fino a che non si ha una netta convergenza verso un valore il più preciso possibile;
- ‘`maxit`’ per stabilire il numero massimo di iterazioni per la massimizzazione;

- ‘`na.action`’ è il metodo per supportare i dati mancanti (per *default* si ha `na.omit`).

Un oggetto creato con `zeroinfl` contiene una lista di componenti che include:

- ‘`stval`’ valori iniziali usati nell’ottimizzazione;
- ‘`par`’ stime di massima verosimiglianza di tutti i parametri;
- ‘`hessian`’ matrice delle derivate seconde della log-verosimiglianza valutata nelle stime di massima verosimiglianza;
- ‘`llh`’ valore del massimo della funzione di log-verosimiglianza;
- ‘`y`’ vettore di dati;
- ‘`x`’ matrice di covariate usata nella parte poissoniana del modello;
- ‘`z`’ matrice di covariate usata nella parte inflazionata di zeri del modello.

Esempio 2

I dati contenuti nel *dataset* ‘*bioChemists*’ riguardano la stesura di articoli scientifici da parte di n = 915 studenti di dottorato in Biochimica. I dati sono disponibili nel sito http://www.indiana.edu/~jslsoc/stata/socdata/couart2.dta nel formato *Stata*.

Si vuole studiare se il numero di articoli scritti dipende da vari fattori che possono influenzare lo studente di dottorato.

In particolare, il *dataset* è così composto:

```
> library(pscl)
> bioChemists
```

	art	fem	mar	kid5	phd	ment
1	0	Men	Married	1	2.520	7
2	0	Women	Single	1	2.050	6
3	0	Women	Single	1	3.750	6
…	…	…	…	…	…	…
913	12	Men	Married	2	1.860	5
914	16	Men	Married	1	1.740	21
915	19	Men	Married	1	1.860	42

Le variabili contenute nel *dataset* sono:

- '`art`' numero di articoli scritti durante gli ultimi 3 anni di dottorato;
- '`fem`' fattore a due livelli, *Men* e *Women* (uomo e donna), che indica il genere dello studente;
- '`mar`' fattore a due livelli che indica lo stato civile dello studente, con i livelli *Single* e *Married* (celibe/nubile e sposato/a);
- '`kid5`' numero di figli con un'età compresa tra 0 e 5 anni;
- '`phd`' prestigio del Dipartimento in cui si svolge l'attività di dottorato;
- '`ment`' numero di articoli redatti dal supervisore del dottorando durante gli ultimi 3 anni.

Si mostra ora la corretta sintassi del comando da impartire per avere una valida stesura della `formula` del modello ZIP. La variabile risposta di tipo conteggio è `art` che può dipendere dalle varie covariate. In questo primo esempio non viene considerata la parte inflazionata di zeri del modello, ma solo la parte relativa alla regressione di Poisson. Nel comando si è espressa la volontà di vedere le iterazioni necessarie all'algoritmo per convergere verso un unico valore, ottenendo in questo modo la stima del massimo della funzione di verosimiglianza (`trace = TRUE`).

```
> data(bioChemists)
> zip <- zeroinfl(count=art ~ .,
         x = ~ fem + mar + kid5 + phd + ment, # no z model
         data=bioChemists,trace=TRUE)

Zero-Inflated Count Model
Using logit to model zero vs non-zero
Using Poisson for counts
dependent variable y:
Y
  0   1   2   3   4   5   6   7   8   9  10  11  12  16  19
275 246 178  84  67  27  17  12   1   2   1   1   2   1   1
generating start values...done
MLE begins...
```

```
initial  value 1665.398520
iter   2 value 1659.747182
iter   3 value 1616.143856
…
iter  19 value 1604.772871
iter  19 value 1604.772871
final  value 1604.772871
converged
done
```

Nel seguente comando, a differenza del primo esempio, viene inserita anche la parte della `formula` riferita al modello con inflazione di zeri. Si indichi con `zip2` la nuova analisi di regressione. I commenti dei comandi sono gli stessi della regressione precedente.

```
> zip2 <- zeroinfl(count=art ~ .,
          + x = ~ fem + mar + kid5 + phd + ment,
          + z = ~ fem + mar + kid5 + phd + ment,
          + data=bioChemists,trace=TRUE)
 ...
```

Gli oggetti `zip` (e `zip2`) contengono diverse quantità:

```
> names(zip)
 [1] "stval" "par" "hessian" "call" "method"  "dist" "llh"
 [8] "y"  "x"  "z"  "xlevels"  "zlevels"  "xTerms"  "zTerms"
[15] "link"    "linkfn"  "n"     "kx"    "kz"
```

Si può accedere agli elementi della lista `zip` attraverso l'operatore $. Ad esempio, il comando `zip$par` restituisce il vettore delle stime dei parametri β_j e γ_r, con $j = 1, \ldots, d_1$ e $r = 1, \ldots, d_2$, d_1 e d_2 numero di parametri nella regressione. Il comando `zip$llh` restituisce il valore del massimo della log-verosimiglianza. Vista l'importanza delle quantità sopra ricordate, esistono alcune funzioni che procedono alla loro estrazione

dall'oggetto `zip` evitando il ricorso all'operatore $. Il comando `coef(zip)` è analogo a `zip$par` e `logLik(zip)` a `zip$llh`. Ad esempio,

```
> zip$llh
[1] -1604.773
```

```
> logLik(zip)
  [1] -1604.773
```

2.3 L'*Output* di `zeroinfl`

Quando si danno le istruzioni per la lettura dei dati relative al modello, è buona norma consultare il `summary`, una generica funzione da usare per poter usufruire delle stime dei coefficienti della regressione. La sintassi generale è semplice ed è la seguente:

```
summary(object, ...)
```

dove per `object` si intende il nome dell'oggetto contenente l'*output* di una analisi di regressione.

L'*output* consiste in una lista dell' analisi della regressione ZIP che include:

- '`Coefficients`' una matrice la cui prima colonna contiene i nomi assegnati a ciascuna covariata, la seconda colonna le stime di massima verosimiglianza di (γ, β), la terza colonna le rispettive stime degli *standard error* e, nella quarta e ultima colonna, i corrispettivi *p-value* calcolati, per *default*, per un sistema di ipotesi di tipo bilaterale H_0: $\theta_j = 0$ *vs* H_1: $\theta_j \neq 0$, dove θ_j è un generico coefficiente del modello;
- '`vc`' la stima della matrice di varianza-covarianza degli stimatori di massima verosimiglianza dei parametri del modello;
- '`beta`' le stime dei coefficienti per la parte poissoniana del modello;
- '`gamma`' le stime dei coefficienti per la parte *zero-inflated* del modello;
- '`theta`' la stima del parametro di sovradispersione, se si adatta ai dati un modello *zero-inflated* binomiale negativo;

- ‘`llh`’ il valore del massimo della log-verosimiglianza sotto le stime correnti dei coefficienti.

Esempio 2 (continuazione)

Essendo la `formula` del modello identica a quella utilizzata nell’*Esempio 2*, si omettono, le iterazioni effettuate dall’algoritmo EM per il calcolo della stima di massima verosimiglianza (`trace = FALSE`). Si inserisca, dunque la `formula` del modello nell’oggetto `zip`:

```
> data(bioChemists)
> zip <- zeroinfl(count=art ~ .,
         + x = ~ fem + mar + kid5 + phd + ment,
         + z = ~ fem + mar + kid5 + phd + ment,
         + data=bioChemists,trace=FALSE)
Zero-Inflated Count Model
Using logit to model zero vs non-zero
Using Poisson for counts
dependent variable y:
Y
  0   1   2   3   4   5   6   7   8   9  10  11  12  16  19
275 246 178  84  67  27  17  12   1   2   1   1   2   1   1
generating start values...done
MLE begins...
done
```

Il modello adattato ai dati può essere riassunto per mezzo del comando `summary`. L’*output* consiste in due parti distinte: la prima riguardante la parte del modello con inflazione di zeri (binomiale) con le relative stime dei coefficienti, mentre la seconda parte si riferisce al modello poissoniano.

```
> summary(zip)
Zero-Inflated Count Model Summary
```

```
--------------------------

Call:
zeroinfl(count = art ~ ., x = ~fem + mar + kid5 + phd + ment,
    z = ~fem + mar + kid5 + phd + ment, data = bioChemists,
trace = T)

Total Log-likelihood: -1604.77287068948

Zero-Inflated Model was fit with a logit link
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -0.794242    0.56519 -1.40526 0.159943
femWomen     0.109760    0.28003  0.39195 0.695092
marMarried  -0.354161    0.31756 -1.11526 0.264739
kid5         0.217138    0.19645  1.10529 0.269035
phd          0.001479    0.14524  0.01018 0.991874
ment        -0.134180    0.04523 -2.96629 0.003014
-----------------------------------------------------------------

Count Model (Poisson)
Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  0.784132   0.132852  5.9023 3.585e-09
femWomen    -0.209164   0.063402 -3.2990 9.703e-04
marMarried   0.103727   0.071108  1.4587 1.446e-01
kid5        -0.143289   0.047428 -3.0212 2.518e-03
phd         -0.006112   0.031007 -0.1971 8.437e-01
ment         0.018084   0.002294  7.8830 3.196e-15
```

La `formula` del modello ZIP adattato è dunque

$$\hat{\eta}_P = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5}$$

$$\hat{\eta}_B = \hat{\gamma}_0 + \hat{\gamma}_1 z_{i1} + \hat{\gamma}_2 z_{i2} + \hat{\gamma}_3 z_{i3} + \hat{\gamma}_4 z_{i4} + \hat{\gamma}_5 z_{i5}\,, \qquad (12)$$

ove $\hat{\eta}_P$ e $\hat{\eta}_B$ indicano i predittori lineari stimati (Bortot, Ventura, Salvan, 2000). Sulla base dell'*output* di `R` si ha

$\hat{\eta}_P = \log(\hat{\lambda})$ = log(numero di `art`)= `0.784132 + (-0.209164) fem + 0.103727 Married + (-0.143289) kid5 + (-0.006112) phd + 0.018084 ment`

e

$\hat{\eta}_B = \log it(p)$ = logit(probabilità 0 `art`) = `-0.794242 + 0.109760 fem + (-0.354161) Married + 0.217138 kid5 + 0.001479 phd + (-0.134180) ment`.

Dunque, esprimendo le precedenti espressioni in funzione dei predittori lineari si ottengono i valori della media di articoli redatti e la probabilità di non scrivere articoli:

$$\hat{\lambda} = \exp(\hat{\eta}_P)$$

e

$$\hat{p} = \frac{\exp(\hat{\eta}_B)}{1+\exp(\hat{\eta}_B)}\,.$$

Risulta di maggior interesse per l'interpretazione del modello, la probabilità di scrivere articoli da parte del dottorando, la quale si può calcolare trovando il complementare a 1 di p, ossia $(1 - p)$.

Si osservi, tuttavia, che nella prima parte le stime dei coefficienti di γ_j non sono tutte significative per la regressione ai livelli fissati $\alpha = 5\%$ e $\alpha = 10\%$. Poiché i *p-value* (α^{oss}) sono maggiori di tali soglie, dato che si accetta ampiamente H_0: $\gamma_j = 0$, volendo approfondire tale regressione si potrebbe eliminare dal modello le covariate che non influenzano la variabile risposta, mantenendo l'intercetta e la variabile `ment` che risulta essere significativa a livello $\alpha = 1\%$.

Nella seconda parte dell'*output*, invece, le stime del parametro β_j sono tutte significative a livello fissato α = 1%, ossia i *p-value* assumono valori < 0.01, per cui si rifiuta l'ipotesi nulla di uguaglianza di β_j a zero in favore dell'ipotesi alternativa.
Il `summary` contiene un riassunto delle variabili presenti in `zip`:

```
> names(summary(zip))
 [1] "call"     "n"     "kx"       "kz"           "dist"
 [6] "gamma"    "beta"  "vc"       "coefficients" "llh"
[11] "method"   "link"
```

In questo caso è necessario creare un oggetto che contenga il summary, dal quale è possibile estrarre le variabili di interesse tramite il simbolo $. Ad esempio sia sommario l'oggetto contenente il summary da cui si vuole estrarre la formula del modello. Allora:

```
> sommario<-summary(zip)
> sommario$call
zeroinfl(count = art ~ ., x = ~fem + mar + kid5 + phd + ment,
    z = ~fem + mar + kid5 + phd + ment, data = bioChemists,
trace = FALSE).
```

2.4 Previsioni sulla Variabile Risposta e Intervalli di Previsione

Quando si modellano dei dati, è interessante scoprire quali valori assumeranno le variabili risposta in relazione alle variabili dipendenti introdotte nel modello. A tale scopo, si introduce la funzione `predict.zeroinfl`, che genera previsioni sulla variabile dipendente in un modello ZIP. La sintassi del comando è

```
predict(object, newdata, se.fit = FALSE, conf = 0.95,
        MC = 1000, type = c("response", "prob"),
        na.action = na.pass, ...)
```

dove:

- ‘`object`’ un oggetto della classe `zeroinfl`;
- ‘`newdata`’ opzionale, un *data frame* nel quale inserire variabili con cui fare le previsioni. Se omesso, sono usati i dati presenti nel *dataset* su cui si fa la regressione;
- ‘`se.fit`’ indica se gli *standard error* dei valori predetti devono essere calcolati;
- ‘`conf`’ la confidenza con cui gli intervalli di previsione sono calcolati, se `se.fit = TRUE`, per *default* è il 95%;
- ‘`MC`’ numero delle iterazioni Monte Carlo (si veda Gentle, 2004) per calcolare gli *standard error* e gli intervalli di confidenza dei valori predetti;
- ‘`type`’ indica il tipo di previsione richiesta. Il *default* è `"response"`, che genera previsioni sulla scala dei dati osservati. Scegliendo l’opzione `type="prob"` si genera una matrice di probabilità predette sul *range* dei dati osservati (analogamente per il comando `predprob.zeroinfl`, si veda Gentle, 2004); non vengono calcolati gli *standard error* o gli intervalli di confidenza per le probabilità predette dal modello. Il valore dell’argomento `type` può essere abbreviato;
- ‘`na.action`’ funzione che determina cosa dovrebbe essere fatto con i valori mancanti nel *newdata*. Il *default* è `NA`;
- ‘`. . .`’ altri argomenti provenienti da altri metodi o da inserire per altri metodi.

Il comando `predict` fornisce una lista di componenti:

- ‘`yhat`’ un vettore di valori predetti dal modello;
- ‘`mu`’ previsioni per la parte poissoniana del modello, sulla scala dei dati;
- ‘`phi`’ le probabilità predette dalla componente con inflazione di zeri del modello, la probabilità che i dati assumano valore pari a zero;
- ‘`se`’ se `se.fit = T`, gli *standard error* dei dati predetti;
- ‘`lower,upper`’ se `se.fit = T`, gli estremi degli intervalli di previsione;
- ‘`prob`’ se `type = "prob"`, una matrice di probabilità predette, con ciascuna riga contenente le probabilità predette in base al *range* dei dati osservati, ad esempio, `ncol(prob) = length(min(y):max(y))`.

Esempio 2 (continuazione)

Si utilizzi il *dataset* '*bioChemists*'. Si adatti il modello ZIP e si indichi con `zip` l'oggetto contenete la regressione. In questo esempio, si desidera calcolare il valore predetto dal modello quando lo studente che scrive articoli scientifici, ad esempio, è un maschio, è sposato, ha un figlio di età inferiore ai 5 anni, il prestigio del Dipartimento è pari a 3.103 e il numero di articoli redatti dal supervisore del dottorando durante gli ultimi 3 anni è pari a 3. Si inseriscano in `R` i primi due comandi dell'*Esempio 2* e si proceda come segue:

```
> newdata <- expand.grid(list(fem="Men",
                              mar="Married",
                              kid5=1,
                              phd=3.103,
                              ment=3))
```

Si può proseguire con la previsione contenuta in `yhat`:

```
> yhat <- predict(zip,newdata=newdata,
                  se.fit=TRUE,MC=2500)
commencing Monte Carlo simulations for predicted counts
MC iterate 1 of 2500
MC iterate 2 of 2500
MC iterate 3 of 2500
…
MC iterate 2498 of 2500
MC iterate 2499 of 2500
MC iterate 2500 of 2500

> yhat
$lower
[1] 1.510928

$upper
```

```
[1] 1.938938

$se
[1] 0.1102761

$yhat
      [,1]
1 1.724670

$mu
      [,1]
1 2.181165

$phi
       [,1]
1 0.2092893
```

Il valore previsto dal modello del numero medie di articoli *E(Y)* (*cfr*. formula (3), Paragrafo 1.2) per uno studente con le caratteristiche presentate all'inizio del paragrafo è `1.72`. Un intervallo di previsione per il valore predetto, a livello fissato $\alpha = 0.05$, è [`1.51`; `1.94`]. Se si desidera fissare $\alpha = 0.01$ è sufficiente aggiungere nel `predict` l'opzione `conf = 0.99`. La probabilità che tale studente non scriva articoli è `0.21` (= `phi`), dunque la probabilità che scriva articoli è (`1 – phi`), ossia `0.79`. La media del numero degli articoli scritti sarà `2.18` (= `mu`). Da non confondere `mu` con `yhat`, in quanto quest'ultimo valore non è nient'altro che il risultato di una ponderazione del numero medio di articoli scritti (`mu`) con la probabilità di produrre articoli (`1 – phi`), ossia in formule si ha `yhat = mu(1 – phi)`e numericamente si ottiene `yhat = 2.18 (1–0.21) = 1.72`, che nell'output di `R` risulta essere approssimato alla sesta cifra decimale.

2.5 Conclusioni

In questo capitolo sono stati introdotti alcuni comandi di R per l'analisi del modello ZIP. In particolare è stata presentata una rassegna sui metodi di adattamento di tale modello, comprendenti i valori previsti dal modello sulla base dei dati a disposizione e corrispettivi intervalli di previsione. Negli esempi svolti è stato preso in considerazione un insieme di dati che riguarda l'ambiente universitario. Nel prossimo capitolo, invece, verranno analizzati con R, dati reali di natura epidemiologica con i metodi illustrati in questo capitolo.

CAPITOLO 3

Applicazione a Dati Reali

3.1 Introduzione

Nei capitoli precedenti è stata presentata una rassegna del modello ZIP dal punto di vista teorico (Capitolo 1) e dal punto di vista di computazionale (Capitolo 2), essendo stati presentati alcuni comandi specifici di `R` per l'adattamento del modello ZIP. In questo capitolo viene discussa un'applicazione ad un *dataset* reale contenente dati di origine medica, in particolare di derivazione epidemiologica, le cui variabili risposta e le cui covariate verranno definite nei paragrafi successivi. Le analisi che sono presentate nel seguito sono state condotte utilizzando il software statistico `R` e la libreria `pscl`.

3.2 I Dati

Il *dataset* è stato fornito dal Dipartimento di Scienze Medico-Diagnostiche e Terapie Speciali dell'Università di Padova e contiene variabili che sono state rilevate per studiare se la quantità di silicio presente nei polmoni, `n.spot`, e il numero di zone del polmone risultate positive al silicio, `n.positive`, dipendono da diverse covariate, qui in seguito riportate e descritte.
Il *dataset* appare nel modo seguente:

	group	Age	Gender	Smoke	n.positive	n.spot
1	exposed	60	M	yes	31	236
2	exposed	49	M	yes	24	184
3	control	66	F	no	3	5
4	control	43	M	yes	4	7
5	control	67	M	yes	4	6
6	control	72	M	no	7	13
7	control	57	M	no	13	14
8	control	65	M	no	5	5
9	control	66	F	no	4	5
10	control	70	M	yes	7	12
11	control	69	F	yes	5	5
12	control	48	M	no	6	18
13	control	76	F	yes	8	12
14	control	72	F	no	5	13
15	control	70	F	yes	0	0
16	control	54	M	yes	0	0
17	control	47	F	yes	0	0
18	control	50	M	no	0	0
19	control	52	F	no	0	0
20	control	66	M	no	0	0
21	control	71	F	yes	0	0
22	control	68	M	no	0	0
23	control	69	M	no	0	0
24	control	49	F	yes	0	0
25	control	77	F	no	0	0
26	control	66	F	no	0	0
27	control	62	M	yes	0	0
28	normal	76	F	no	6	6
29	normal	65	M	yes	4	8
30	normal	74	M	no	6	11
31	normal	50	F	no	0	0
32	normal	82	F	no	0	0
33	normal	70	M	no	0	0

```
34  normal  16      F    no          0      0
35  normal  61      M   yes          0      0
36  normal  82      F    no          0      0
37  normal  66      M    no          0      0
38  normal  61      M    no          0      0
39  normal  48      F    no          0      0
40  normal  68      M    no          0      0
41  normal  75      M    no          0      0
42  normal  68      M   yes          0      0
43  normal  80      M    no          0      0
44  normal  47      M   yes          0      0
45  normal  67      M    no          0      0
46  normal  38      F   yes          0      0
47  normal  55      F    no          0      0
```

Si dispone, dunque, di osservazioni su $n = 47$ pazienti, sui quali sono state misurate le seguenti variabili:

- ‘`group`’: fattore a tre livelli (`exposed`, `control`, `normal`) indicante le caratteristiche del paziente: decesso per tumore al polmone ed esposizione nel lavoro al silicio (`exposed`), decesso per tumore al polmone per cause ignote (`control`), decesso per altre ragioni (`normal`);
- ‘`Age`’: età dei soggetti considerati;
- ‘`Gender`’: fattore a due livelli, `M` e `F` (maschio e femmina), che indica il genere del soggetto;
- ‘`Smoke`’: fattore a due livelli, `yes` e `no`, che segna se il soggetto era fumatore oppure non fumatore;
- ‘`n.positive`’: variabile che determina il numero di zone del polmone ritenute positive, in quanto contenti tracce di silicio. Tale variabile assume valori compresi tra `0` e `64`;
- ‘`n.spot`’: variabile che indica la quantità di silicio trovata nelle aree di suddivisione del polmone ritenute positive. Tale variabile assume potenzialmente valori compresi tra `0` e $+\infty$.

3.3 Analisi Preliminari

Con l'analisi preliminare si intende effettuare una prima analisi esplorativa, anche attraverso l'utilizzo di opportuni grafici, quali ad esempio il *boxplot*.

Iniziando l'analisi dalla variabile `group`, si nota che il valore '`control`' è assunto, approssimativamente, dal `53.19` % degli individui, il valore '`exposed`' riguarda soltanto il `4.25` % degli individui; il valore '`normal`' è assunto dal `42.56` % degli individui osservati. Dato che soltanto 2 osservazioni per la variabile `group` assumono valore `exposed` (ossia il `4.25` % del totale), queste verranno escluse dall'analisi poiché non contengono sufficienti informazioni per essere considerate nello studio dei dati. A supporto di questa affermazione, si può notare dalla Figura 1, sia per quanto riguarda la variabile `n.spot` e sia per `n.positive`, che sono infatti presenti due valori troppo estremi per essere inclusi nell'analisi che verrà condotta nei paragrafi successivi. Una volta stimato il modello, si proverà ad ottenere questi valori mediante il processo di previsione.

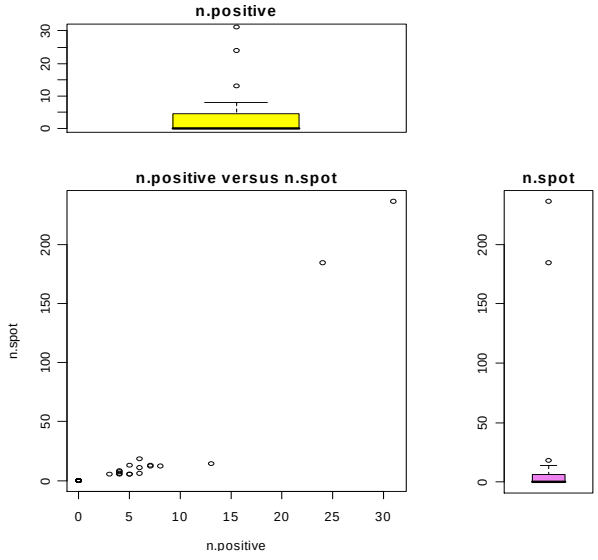

Figura 1: Grafico di dispersione[1] di `n.spot` *versus* `n.positive`.

[1] Per la costruzione del grafico di dispersione si sono eseguiti i seguenti comandi in R:

```
par(fig=c(0.0,0.7,0.0,0.7), mar=c(4,4,2,2)); plot(n.positive,n.spot);
par(fig=c(0.0,0.7,0.0,0.7), mar=c(4,4,2,2)); plot(n.positive,n.spot);
title('n.positive versus n.spot'); par(fig=c(0.0,0.7,0.7,1), mar=c(2,4,2,2),
new=T);boxplot(n.positive, main='n.positive', col='yellow');
par(fig=c(0.7,1,0.05,0.7), mar=c(2,4,2,2),new=T);
boxplot(n.spot, main='n.spot', col='violet')
```

Le osservazioni relative al gruppo di controllo superano il `57` %. Per quanto riguarda la suddivisione per genere, il `42.55` % sono femmine, il `57.45` % sono maschi. I fumatori rappresentano il `38.30` % e i non fumatori il `61.70` %. Suddividendo le osservazioni per genere e per abitudine al fumo si ottengono i seguenti risultati:

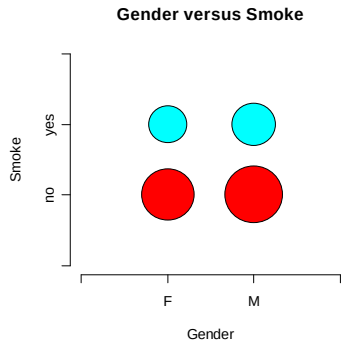

Figura 2: Diagramma a bolle[2] della variabile `Gender` suddivisa per `Smoke`.

Dalla Figura 2 si deduce che la maggior concentrazione si ha per i maschi non fumatori, mentre la minore concentrazione si ha per le femmine fumatrici. Il test χ^2 di Pearson[3] per verificare se le due variabili sono indipendenti, porta ad accettare l'ipotesi di indipendenza a livello 5 % (`X-squared = 0.0594, p-value = 0.8074`).

Per quanto riguarda le variabili `Age`, `n.positive` e `n.spot`, sono stati disegnati i rispettivi *boxplot* dai quali si deducono le mediane, i quartili, il *range*, eventuali valori anomali e le caratteristiche della forma della distribuzione.

La variabile `Age` ha un range che va da poco meno di 40 a poco più di 80, pur presentando un valore anomalo pari a 16. La relativa distribuzione assume una lieve asimmetria negativa.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.00   53.00   66.00   62.34   70.00   82.00
```

[2] Il grafico a bolle presentato prende il nome di '`bubbleplot`' ed è disponibile installando in `R` il pacchetto statistico `labstatR` tramite il comando `library(labstatR)`.

[3] In `R`: `chisq.test(x,y)`, con `x` contenente un vettore o una matrice, `y` contenente un vettore ma ignorato se `x` è una matrice.

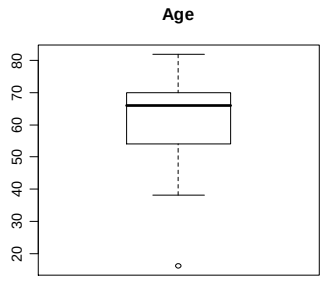

Figura 3: *Boxplot* della variabile `Age`.

Per quanto riguarda la variabile `n.positive`, la maggior parte delle osservazioni assume valore nullo, più precisamente il 67 % delle unità statistiche. La distribuzione, proprio per questa particolarità, risulta asimmetrica rispetto alla mediana che non coincide assolutamente con la media ma che coincide con il primo quartile (Figura 4).

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   0.000   0.000   1.933   4.000  13.000
```

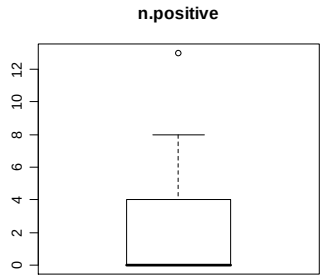

Figura 4: *Boxplot* della variabile `n.positive`.

Per quanto riguarda la variabile `n.spot`, si può dire che il suo *range* è piuttosto ristretto dato che il primo quartile e la mediana sono pari a zero; ciò è dovuto alla presenza prevalente di zeri nei dati (in percentuale 67%). La distribuzione è fortemente asimmetrica (Figura 5). Anche in questo caso i valori dei due individui esposti sono stati esclusi dalla rappresentazione grafica (Figura 5).

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   0.000   0.000   3.111   5.000  18.000
```

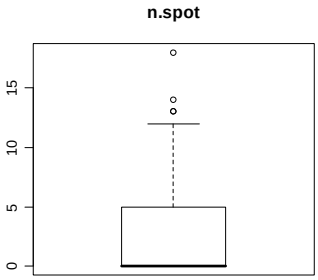

Figura 5: *Boxplot* della variabile `n.spot`.

Le distribuzioni delle variabili `n.positive` e `n.spot` sono ben visibili dai grafici sottostanti dai quali si può meglio notare che parecchie osservazioni assumono valore pari a zero. Questa caratteristica giustifica la scelta di adattare un modello che preveda questo squilibrio, dato l'eccesso di zeri nelle due variabili.

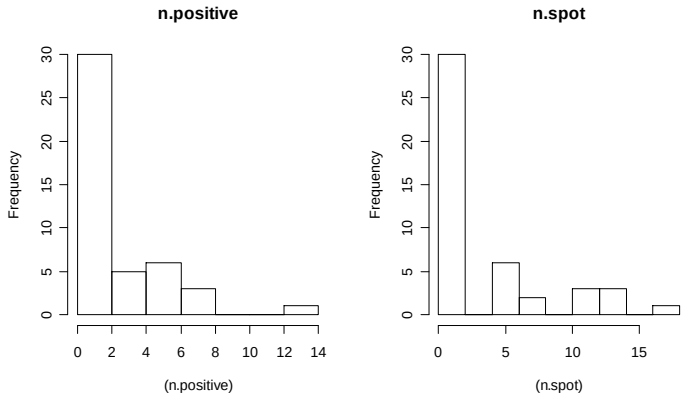

Figura 6: Istogramma delle variabili risposta `n.positive` e `n.spot`.

3.4 Le Variabili `n.spot` e `n.positive`

Si desidera ora studiare come si distribuiscono le due variabili risposta, rispetto alle diverse covariate. Si inizi dalla variabile `group` e si prosegua la stessa analisi per le altre variabili.

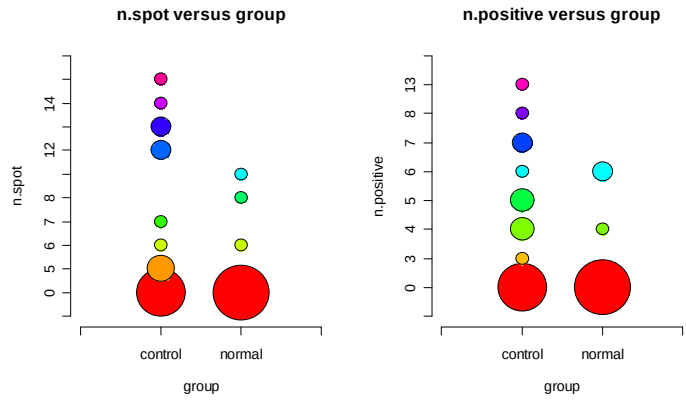

Figura 7: Diagramma a bolle di `n.spot` e `n.positive` suddivise per `group`.

Dalla Figura 7 si evince che, per la maggior parte delle osservazioni, non sono presenti delle quantità significative di `spot` di silicio e non è stato rilevato un numero considerevole di zone del polmone positive al silicio, per i soggetti appartenenti sia al gruppo di controllo sia nel gruppo di soggetti deceduti per cause naturali. Ciò fa pensare che non ci sia una relazione tra la quantità di silicio trovata nei polmoni e il tipo di decesso. Inoltre, dai *boxplot* della Figura 8, si può notare che in termini di *range* vi è una forte diversità tra i due gruppi ma, per entrambi i gruppi, il primo quartile e la mediana coincidono. I risultati ottenuti sono molto simili per `n.spot` e `n.positive` ed, infatti, è ragionevole asserire che le due variabili sono strettamente correlate (la correlazione vale 0.92). Per verificare se le mediane coincidono, ad esempio per provare che per `n.spot` il gruppo di controllo e il gruppo normale hanno la stessa mediana e analogamente per `n.positive`, si usa il test Kruskal[4], ossia un test non parametrico. In definitiva, il test porta a rifiutare l'ipotesi di uguaglianza in mediana della variabile `group`, sia per quanto riguarda `n.spot` (`Kruskal-Wallis chi-squared = 5.2365, p-value = 0.02212`) sia per quanto riguarda `n.positive` (`Kruskal-Wallis chi-squared = 5.1154, p-value = 0.02371`), in quanto i *p-value* sono minori del 5 %. La percentuale di zeri in `n.spot` e `n.positive` è 52 % per i pazienti deceduti per tumore, mentre per i pazienti morti per cause naturali la percentuale vale 85 %.

[4] In `R` i comandi sono rispettivamente: `kruskal.test(n.spot~group)` e `kruskal.test(n.positive~group)`.

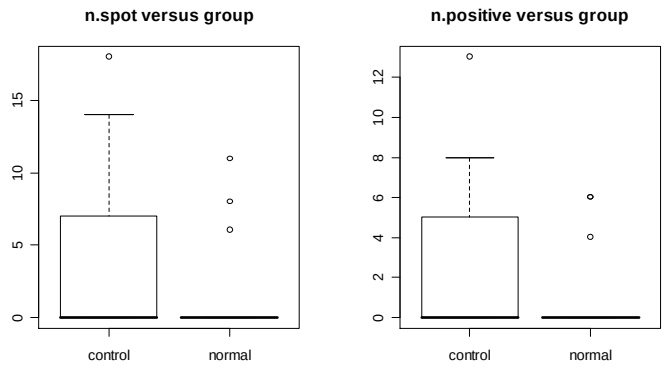

Figura 8: *Boxplot* di `n.spot` e `n.positive` suddivise per le modalità della variabile `group`.

Si considera ora la variabile età (`Age`). Dalla Figura 9, in generale, si può dedurre che `n.positive`, e anche `n.spot`, assumono spesso il valore zero indipendentemente dall'età. La correlazione tra `n.spot` e `Age` vale `0.1152974`, mentre tra `n.positive` e `Age` vale `0.1523107`.

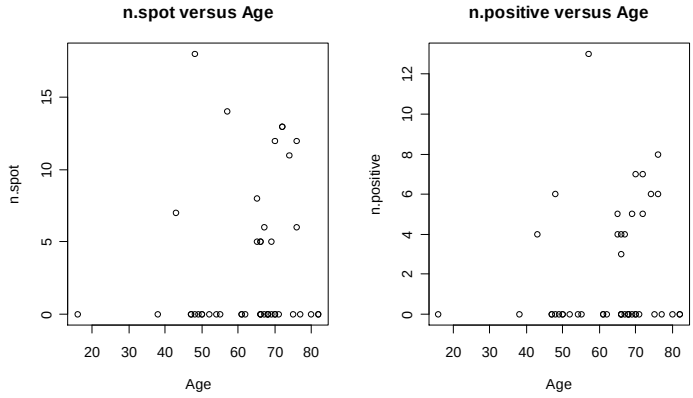

Figura 9: Grafico di dispersione di `n.spot` e `n.positive` per `Age`.

Si passa ora alla variabile genere (`Gender`). Dalla Figura 10 si nota che il numero di zeri nella variabile risposta è parecchio elevato sia per i maschi che per le femmine in quanto primo quartile e mediana coincidono e assumono valore pari a zero; i *boxplot* sono alquanto simili per posizione nella scala dei dati e per dimensione della scatola. Per stabilire se il livello di `n.spot` e di `n.positive` è uguale in mediana rispetto al genere, si conduce il test di Kruskal, il quale fa accettare tale ipotesi a livello fissato α

= 5 % e 10 % per entrambe le variabili risposta (per `n.spot` `Kruskal-Wallis chi-squared` = `0.541`, `p-value` = `0.462`, per `n.positive` `Kruskal-Wallis chi-squared` = `0.3121`, `p-value` = `0.5764`).

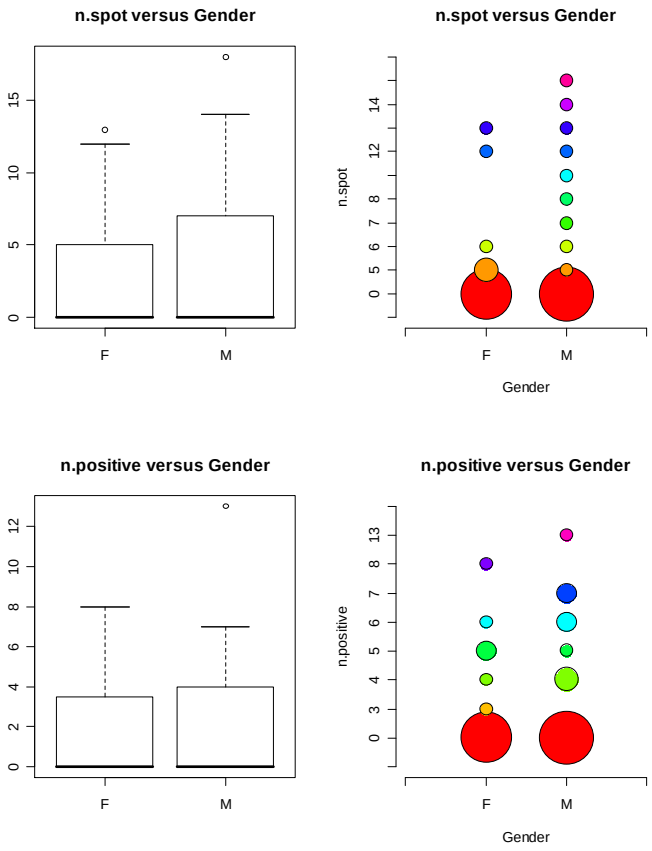

Figura 10: `n.spot` e `n.positive` suddivise per genere mediante *boxplot* e diagramma a bolle.

Osservando i *boxplot* delle variabili `n.spot` e `n.positive` rispetto all'abitudine al fumo (Figura 11), non si evince una sostanziale disuguaglianza tra i fumatori e i non fumatori. In mediana, i fumatori e i non fumatori coincidono, sia per la variabile `n.spot` (`Kruskal-Wallis chi-squared` = `0.0968`, `p-value` = `0.7557`) sia per la variabile `n.positive` (`Kruskal-Wallis chi-squared` = `0.0968, df` = `1, p-value` = `0.7557`) a livello fissato α = 5 % e 10 %.

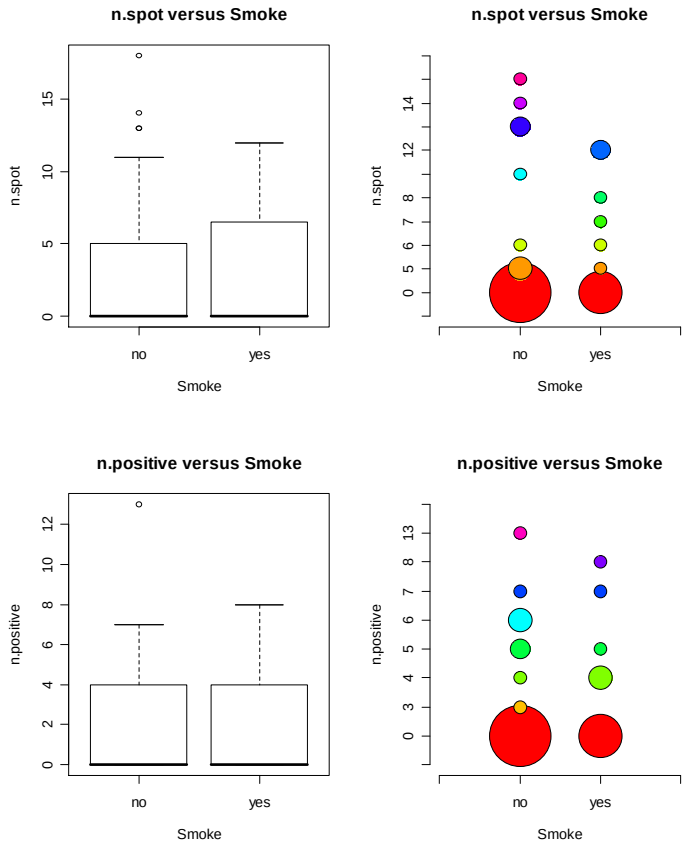

Figura 11: `n.spot` e `n.positive` suddivise per soggetti fumatori e non fumatori.

Si considera, infine, `n.positive` *versus* `n.spot`. Dai due diagrammi di dispersione (Figura 12) è possibile notare un notevole affollamento di zeri nella variabile `n.spot`, come è ragionevole supporre, prevalentemente in presenza di `n.positive = 0`. Infatti, dove non vi sono zone risultate positive alla presenza di silicio, è ovvio che non è possibile rilevare alcuna dose di silicio. La correlazione, infatti, è elevata e assume un valore prossimo a `1` (`0.92`).

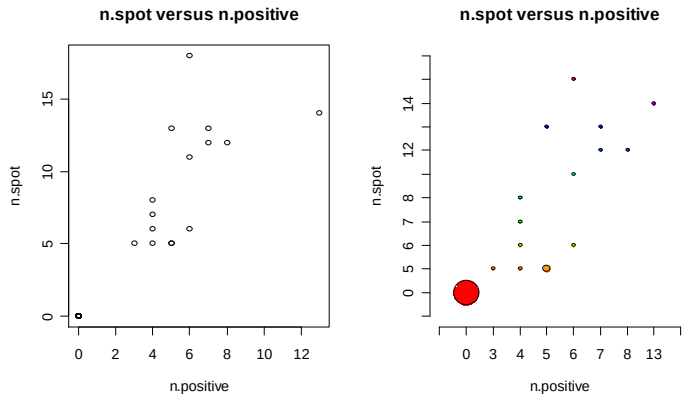

Figura 12: `n.spot` raffigurato per `n.positive`.

Concludendo, sembra esistere una lieve relazione tra le variabili `n.spot` e `n.positive` e la modalità di decesso. Per quanto riguarda le variabili `Age`, `Smoke` e `Gender`, queste non sembrano influire sulle variabili risposta `n.positive` e `n.spot`.

3.5 Interazioni tra le Variabili

L'effetto di interazione è valutabile visivamente attraverso particolari grafici, detti `interaction plot`. Si intende valutare se due variabili si possano influenzare vicendevolmente. Si inizi l'analisi tenendo in considerazione la variabile risposta `n.spot` (Figura 13).

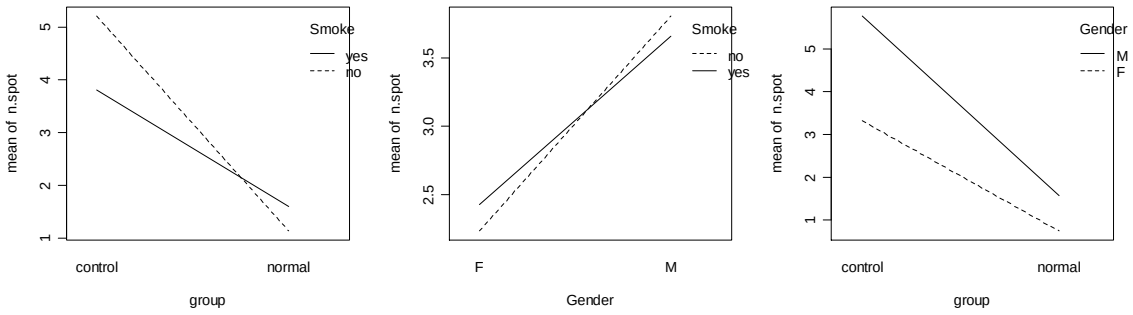

Grafico 1 Grafico 2 Grafico 3

Figura 13: *Interaction plots*.

Per quanto riguarda il Grafico 1 (Figura 13), si nota che la media di `n.spot` è più alta per i non fumatori che sono deceduti a causa del cancro ai polmoni, mentre per coloro, comunque non fumatori, che sono morti per cause naturali, è stata rilevata in media una minore quantità di `n.spot`. Per i fumatori, inoltre, si verifica lo stesso evento, ossia per i fumatori appartenenti al gruppo di controllo la media di n.spot è più alta rispetto alla media per i fumatori morti naturalmente. Tuttavia, sembra esserci un'interazione tra le variabili `group` e `Smoke`. Il Grafico 2 (Figura 13) manifesta una possibile interazione tra le variabili `Gender` e `Smoke`. La media di *spot* di silicio è più alta per i maschi rispetto alle femmine, indipendentemente dall'attitudine al fumo; per le femmine la media di n.spot tende a essere nulla per le non fumatrici, mentre per le fumatrici la media è diversa da zero ma raggiunge un valore relativamente piuttosto basso. Il Grafico 3 (Figura 13), come è intuitivo supporre, conferma l'inesistenza di un'eventuale interazione tra le variabili `group` e `Gender`.

Si passa ora alla variabile `n.positive` e all'approfondimento delle relazioni tra le variabili (Figura 14).

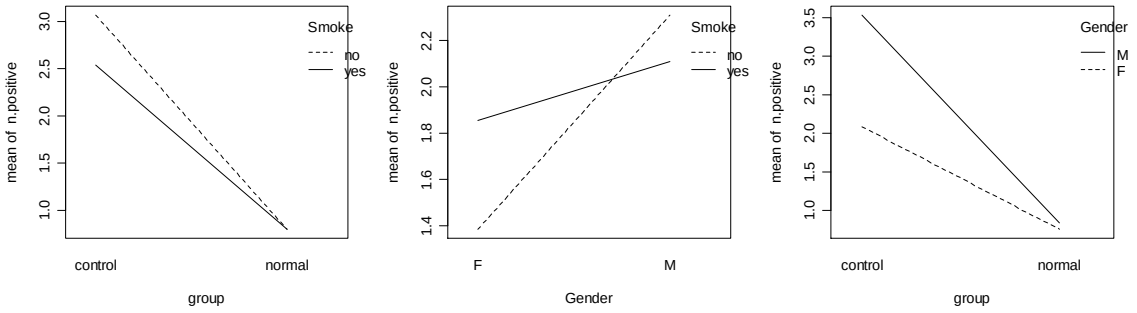

Grafico 1 Grafico 2 Grafico 3

Figura 14: *Interaction plots*.

Il Grafico 2 (Figura 14) mostra che la media di zone polmonari positive alla presenza di silicio è, per i fumatori, significativamente più alta nei soggetti deceduti per cancro rispetto ai soggetti comunque fumatori, ma deceduti per cause naturali. Anche per i non fumatori, considerando la modalità di decesso, si evince una discreta differenza in termini di media di `n.positive`. È probabile che possa esistere un'interazione tra le variabili `group` e `Smoke`. Nel Grafico 2 (Figura 14), la media di zone positive

(`n.positive`) è più elevata per i soggetti maschi non fumatori, mentre per i soggetti femmina non fumatori tale media è piuttosto bassa e si discosta di parecchio dalla media dei soggetti femmina fumatori. Osservando tale grafico, si può intuire che è presente un'interazione tra `Gender` e `Smoke`. Il Grafico 3 (Figura 14) mostra un'assenza di interazione tra le variabili `group` e `Gender`, come è ragionevole aspettarsi: per la modalità normal la media di zone positive è bassa per entrambi i generi, per la modalità control, al contrario, la media è significativamente diversa da zero sia per i maschi sia per le femmine. Dopo queste analisi, si può affermare, dunque, che potrebbe essere presente l'effetto `Smoke`.

3.6 Adattamento del Modello

Partendo dalle analisi preliminari, si è notato un sovraffollamento di zeri per la variabile `n.spot`, strettamente collegata, ovviamente, alla variabile `n.positive`. Questo eccesso di zeri nella variabile risposta è trattabile mediante il modello di Poisson con inflazione di zeri, ossia il modello ZIP presentato nel Capitolo 1. Si passi, dunque, all'adattamento di tale modello in R per mezzo dei comandi presentati nel Capitolo 2. Successivamente, si desidera valutare quali covariate influenzano le variabili risposta n.spot e `n.`positive considerando due modelli separatamente. Inizialmente si partirà da un modello, dapprima per `n.spot`, comprendente le interazioni tra le covariate trovate nelle precedenti analisi e, successivamente, si condurrà un'analisi *backward* per snellire i modelli. Le variabili o le interazioni saranno tolte una alla volta stabilendo gradualmente se la preferenza viene data a un modello più parsimonioso o a un modello più complesso.

3.6.1 Il Modello con Variabile Risposta `n.spot`

Si consideri il modello ZIP contenente tutte le covariate del *dataset* in questione e si consideri la variabile risposta `n.spot`. Si inizi la modellazione dei dati inserendo nella

regressione le variabili esplicative e le interazioni a coppie tra di esse. La formula del modello, dunque, è la seguente:

```
zip<-zeroinfl(count = n.spot~.,
x = ~ group+Age + Gender + Smoke + group:Age + group:Gender +
group:Smoke + Age:Gender + Age:Smoke + Gender:Smoke,
z = ~ group + Age + Gender + Smoke + group:Age + group:Gender +
group:Smoke + Age:Gender + Age:Smoke + Gender:Smoke,
trace=F, data=dati)
```

Il massimo della funzione di log-verosimiglianza è `-55.94`. Le stime dei coefficienti dei parametri sono:

```
Zero-Inflated Model was fit with a logit link
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -12.22538    24.0912 -0.5075   0.6118
groupnormal            6.59498    17.1449  0.3847   0.7005
Age                    0.16971     0.3435  0.4940   0.6213
GenderM                9.37032    25.2348  0.3713   0.7104
Smokeyes               7.28950     8.7715  0.8310   0.4059
groupnormal:Age       -0.09018     0.1941 -0.4645   0.6423
groupnormal:GenderM    1.32305     5.1240  0.2582   0.7962
groupnormal:Smokeyes  -0.58724     2.4918 -0.2357   0.8137
Age:GenderM           -0.12363     0.3629 -0.3407   0.7333
Age:Smokeyes          -0.10062     0.1202 -0.8371   0.4025
GenderM:Smokeyes      -1.63387     2.4110 -0.6777   0.4980
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                    Estimate Std. Error z value  Pr(>|z|)
(Intercept)         -9.03877    3.19823 -2.8262 4.711e-03
```

```
groupnormal           -5.23535   14.68658 -0.3565 7.215e-01
Age                    0.16085    0.04609  3.4896 4.837e-04
GenderM               12.97825    3.22737  4.0213 5.787e-05
Smokeyes              -3.11275    1.69874 -1.8324 6.689e-02
groupnormal:Age        0.05037    0.19410  0.2595 7.952e-01
groupnormal:GenderM    1.72384    0.74420  2.3164 2.054e-02
groupnormal:Smokeyes   0.15462    1.82553  0.0847 9.325e-01
Age:GenderM           -0.18459    0.04693 -3.9332 8.381e-05
Age:Smokeyes           0.03315    0.02270  1.4606 1.441e-01
GenderM:Smokeyes       0.72334    0.55467  1.3041 1.922e-01
```

Per la parte del modello con inflazione di zeri non risulta significativo alcun coefficiente a livello fissato $\alpha = 5\%$. Per la parte poissoniana del modello, al contrario, i coefficienti risultano tutti significativi al livello $\alpha = 1\ \%$.

Si provi a semplificare la regressione togliendo, una alla volta, le variabili che non influiscono sulla variabile risposta per la parte con inflazione di zeri. Si inizi togliendo, dapprima, l'interazione tra `group` e `Smoke` (*p-value* = `0.8137`). Si mantenga, naturalmente, invariata la parte di modellazione di Poisson. Il valore massimo assunto dalla funzione di log-verosimiglianza, in questo caso, è `-55.94`. Si osservi l'output:

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -12.97329    18.1171 -0.7161   0.4739
groupnormal           5.19853    12.6492  0.4110   0.6811
Age                   0.18033     0.2583  0.6982   0.4851
GenderM              10.10839    18.8065  0.5375   0.5909
Smokeyes              7.58604     8.2934  0.9147   0.3603
groupnormal:Age      -0.07395     0.1529 -0.4836   0.6287
groupnormal:GenderM   1.37342     3.8297  0.3586   0.7199
Age:GenderM          -0.13314     0.2706 -0.4920   0.6227
Age:Smokeyes         -0.10503     0.1138 -0.9231   0.3560
GenderM:Smokeyes     -1.82706     2.1450 -0.8518   0.3943
```

```
------------------------------------------------------------
```

```
Count Model (Poisson)
Coefficients:
                     Estimate Std. Error z value  Pr(>|z|)
(Intercept)          -9.02838    3.12927 -2.8851 3.912e-03
groupnormal          -5.82434   13.02654 -0.4471 6.548e-01
Age                   0.16071    0.04512  3.5618 3.684e-04
GenderM              12.97114    3.15576  4.1103 3.951e-05
Smokeyes             -3.11826    1.69458 -1.8401 6.575e-02
groupnormal:Age       0.05814    0.17280  0.3365 7.365e-01
groupnormal:GenderM   1.73768    0.70474  2.4657 1.367e-02
groupnormal:Smokeyes  0.22409    1.64754  0.1360 8.918e-01
Age:GenderM          -0.18450    0.04592 -4.0176 5.879e-05
Age:Smokeyes          0.03323    0.02268  1.4656 1.428e-01
GenderM:Smokeyes      0.72338    0.55105  1.3127 1.893e-01
```

A livello fissato α = 5 %, per la prima parte del modello, non risulta significativo nessun coefficiente. Contrariamente, per la parte poissoniana valgono le osservazioni rilevate nell'adattamento precedente. Si prosegua nello sfoltimento dei parametri per quanto riguarda la parte del modello con inflazione di zeri, togliendo l'interazione tra `group` e `Gender` (`1.37342` con *p-value* `0.7199`). In questo caso la stima del massimo della log-verosimiglianza è `-56.06`.

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.39350    16.4888 -0.4484   0.6539
groupnormal       7.80146     9.1817  0.8497   0.3955
Age               0.09998     0.2289  0.4367   0.6623
GenderM           4.50407    15.2522  0.2953   0.7678
Smokeyes          7.05612     8.0025  0.8817   0.3779
groupnormal:Age  -0.09375     0.1339 -0.7003   0.4837
```

```
Age:GenderM       -0.05150     0.2071 -0.2487   0.8036
Age:Smokeyes      -0.09523     0.1112 -0.8567   0.3916
GenderM:Smokeyes  -1.92449     1.9633 -0.9802   0.3270
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                     Estimate Std. Error  z value  Pr(>|z|)
(Intercept)          -8.62891    3.48018 -2.47944 0.0131589
groupnormal          -4.63794   20.27732 -0.22873 0.8190824
Age                   0.15499    0.05013  3.09167 0.0019904
GenderM              12.56610    3.51026  3.57982 0.0003438
Smokeyes             -3.08113    1.70620 -1.80584 0.0709432
groupnormal:Age       0.04306    0.26708  0.16122 0.8719170
groupnormal:GenderM   1.66608    0.84928  1.96176 0.0497908
groupnormal:Smokeyes  0.09174    2.46344  0.03724 0.9702919
Age:GenderM          -0.17869    0.05096 -3.50661 0.0004539
Age:Smokeyes          0.03313    0.02274  1.45680 0.1451711
GenderM:Smokeyes      0.69101    0.56456  1.22398 0.2209613
```

Togliendo l'interazione nella parte del modello con inflazione di zeri, i parametri continuano ad essere tutti non significativi al livello fissato α = 5 % e 10 %. Inoltre, alcune stime dei parametri della parte poissoniana risultano poco rilevanti per la regressione e, in particolare, il coefficiente che risulta meno significativo si riferisce all'interazione tra `group` e `Smoke`. Si provi a togliere le interazioni tra `Age` e `Gender` e tra `group` e `Smoke`, per la prima e la seconda parte del modello, rispettivamente. Si ottiene il seguente output:

```
Total Log-likelihood: -56.1236315214307

 Zero-Inflated Model was fit with a logit link
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)       -3.41861    5.59597 -0.6109   0.5413
groupnormal        6.23895    7.37714  0.8457   0.3977
Age                0.04455    0.07953  0.5602   0.5754
GenderM            0.73554    1.13324  0.6491   0.5163
Smokeyes           6.93838    7.75755  0.8944   0.3711
groupnormal:Age   -0.06970    0.10617 -0.6565   0.5115
Age:Smokeyes      -0.09368    0.10826 -0.8653   0.3869
GenderM:Smokeyes  -1.86104    1.89467 -0.9823   0.3260
---------------------------------------------------------

Count Model (Poisson)
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)        -8.22037    3.37869 -2.4330 0.0149741
groupnormal        -3.86720    4.81414 -0.8033 0.4218005
Age                 0.14917    0.04872  3.0620 0.0021986
GenderM            12.16467    3.39082  3.5875 0.0003338
Smokeyes           -3.05557    1.70306 -1.7942 0.0727876
groupnormal:Age     0.03334    0.06336  0.5263 0.5986902
groupnormal:GenderM 1.61765    0.73223  2.2092 0.0271600
Age:GenderM        -0.17299    0.04923 -3.5141 0.0004413
Age:Smokeyes        0.03317    0.02280  1.4548 0.1457263
GenderM:Smokeyes    0.66527    0.55448  1.1998 0.2302187
```

Per valutare se risulta migliore il modello più complicato o quello più semplice si usa Test Rapporto di Verosimiglianza (*TRV*),

$$TRV = 2\left(l\left(H_1\right) - l\left(H_0\right)\right) \sim \chi^2_{p-p_0} \text{ sotto } H_0.$$

Per H_0 si intende il modello più semplice più parsimonioso, mentre H_1 sottointende il modello più complicato con più parametri; p e p_0 sono il numero di parametri del modello più complicato e più semplice, rispettivamente. Il *TRV*, che deve essere

confrontato con un χ^2_2, fornisce una netta evidenza in favore del modello con meno parametri:

```
> 1-pchisq(2*(-56.0630346985277-(-56.1236315214307)), 20-18)
[1] 0.9412026
```

Nell'adattamento successivo si provi a migliorare il modello precedente togliendo la variabile `Age` e l'interazione tra `group` e `Age`. La stima di massima verosimiglianza ottenuta è `-56.29`.

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.43164    5.55893 -0.6173   0.5370
groupnormal       6.51447    7.28623  0.8941   0.3713
GenderM           0.71524    1.11562  0.6411   0.5214
Smokeyes          7.24698    7.76567  0.9332   0.3507
groupcontrol:Age  0.04495    0.07920  0.5675   0.5704
groupnormal:Age  -0.02837    0.09421 -0.3011   0.7633
Smokeyes:Age     -0.09810    0.10854 -0.9038   0.3661
GenderM:Smokeyes -1.90861    1.89530 -1.0070   0.3139
-------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)         -8.33956    3.35454  -2.486 0.0129169
groupnormal         -1.34599    0.56310  -2.390 0.0168329
Age                  0.15086    0.04835   3.120 0.0018069
GenderM             12.10237    3.38237   3.578 0.0003461
Smokeyes            -2.74074    1.60509  -1.708 0.0877239
groupnormal:GenderM  1.40832    0.62833   2.241 0.0250022
Age:GenderM         -0.17130    0.04905  -3.492 0.0004792
```

```
Age:Smokeyes          0.02883    0.02137   1.349 0.1772672
GenderM:Smokeyes      0.55419    0.51556   1.075 0.2824018
```

Dalla precedente analisi, per quanto concerne la prima parte del modello, tutti i parametri sembrano non essere significativi, mentre per la seconda parte del modello ci sono variabili e interazioni che hanno un'influenza sulla variabile risposta. Si tolgano da quest'ultimo adattamento le variabili con *p-value* più alti e, dunque, poco importanti nella spiegazione della variabile risposta. In questo caso si tratta dell'interazione tra `group` e `Age` nella prima parte e tra `Gender` e `Smoke` nella seconda. L'*output* è il seguente:

```
Total Log-likelihood: -57.2695465883864

 Zero-Inflated Model was fit with a logit link
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.12559    4.81607 -0.2337  0.81521
groupnormal       1.59265    0.90528  1.7593  0.07853
GenderM           0.59124    1.12216  0.5269  0.59828
Smokeyes          5.82894    7.08322  0.8229  0.41055
Smokeno:Age       0.01004    0.06667  0.1506  0.88032
Smokeyes:Age     -0.06502    0.07769 -0.8370  0.40261
GenderM:Smokeyes -1.84971    1.83366 -1.0088  0.31309
-------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)        -7.02258    3.18877  -2.202 0.0276452
groupnormal        -1.12417    0.52535  -2.140 0.0323663
Age                 0.13041    0.04541   2.872 0.0040808
GenderM            10.58897    3.12538   3.388 0.0007039
Smokeyes           -1.66829    1.24850  -1.336 0.1814706
```

```
groupnormal:GenderM  1.17961    0.59216   1.992 0.0463653
Age:GenderM         -0.14707    0.04427  -3.322 0.0008943
Age:Smokeyes         0.01877    0.01921   0.977 0.3285776
```

Il confronto del *TRV* con un χ^2_2 porta all'accettazione del modello più semplice che include meno parametri (`0.3745062` > 0.05 = α livello fissato).

```
> 1-pchisq(2*(-56.2873995493395-(-57.2695465883864 )),17-15)
[1] 0.3745062
```

Si semplifichi il modello precedente escludendo dall'analisi le interazioni tra `Smoke` e Age in entrambe le parti del modello. La stima di massima log-verosimiglianza vale `-58.14`. Le stime dei parametri sono:

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.4533     0.8088 -0.5604  0.57519
groupnormal        1.5908     0.8084  1.9679  0.04908
GenderM            0.5675     0.9536  0.5952  0.55174
Smokeyes           0.5458     1.2809  0.4261  0.67004
GenderM:Smokeyes  -1.0983     1.5837 -0.6935  0.48800
-----------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                    Estimate Std. Error z value  Pr(>|z|)
(Intercept)          -8.3762    2.67636  -3.130 1.750e-03
groupnormal          -1.2947    0.48472  -2.671 7.561e-03
Age                   0.1506    0.03773   3.992 6.558e-05
GenderM              11.4389    2.79936   4.086 4.384e-05
Smokeyes             -0.4721    0.19113  -2.470 1.352e-02
groupnormal:GenderM   1.2974    0.57000   2.276 2.284e-02
```

```
Age:GenderM          -0.1592    0.03973  -4.007 6.138e-05
```

La parte del modello con inflazione di zeri mantiene ancora alcuni coefficienti poco significativi per la variabile risposta. La parte poissoniana, al contrario, contiene tutti parametri significativi per `n.spot`. Si modifichi l'adattamento precedente ai dati eliminando la variabile Smoke e lasciando, naturalmente, inalterata la seconda parte. L'*output* della nuova modellazione risulta essere il seguente:

```
Total Log-likelihood: -58.138270863291

 Zero-Inflated Model was fit with a logit link
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.4575     0.8092 -0.5654  0.57183
groupnormal       1.5928     0.8085  1.9700  0.04884
GenderM           0.5720     0.9539  0.5997  0.54871
GenderF:Smokeyes  0.5537     1.2811  0.4322  0.66561
GenderM:Smokeyes -0.5551     0.9228 -0.6015  0.54750
-------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)        -8.3759    2.67625  -3.130 1.750e-03
groupnormal        -1.2943    0.48466  -2.671 7.572e-03
Age                 0.1506    0.03773   3.992 6.558e-05
GenderM            11.4393    2.79929   4.087 4.379e-05
Smokeyes           -0.4719    0.19114  -2.469 1.356e-02
groupnormal:GenderM  1.2973    0.56997   2.276 2.284e-02
Age:GenderM         -0.1592    0.03973  -4.008 6.127e-05
```

La prima parte del modello mantiene variabili non significative, la seconda parte, invece, continua a ritenere significative le stesse variabili esplicative dell'adattamento

ai dati stimato precedentemente. Si tolgano le variabili non significative per la prima parte del modello, in particolare l'interazione tra `Gender` e `Smoke`. `R` fornisce ulteriori stime:

```
Total Log-likelihood: -58.4117897249606

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2424     0.6309 -0.3843  0.70079
groupnormal   1.5173     0.7856  1.9314  0.05343
GenderM       0.1750     0.7652  0.2288  0.81906
-------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)         -8.3875    2.66346  -3.149 1.638e-03
groupnormal         -1.2946    0.48460  -2.672 7.551e-03
Age                  0.1508    0.03755   4.016 5.925e-05
GenderM             11.4524    2.78618   4.110 3.949e-05
Smokeyes            -0.4738    0.19080  -2.483 1.303e-02
groupnormal:GenderM  1.2982    0.56979   2.278 2.270e-02
Age:GenderM         -0.1594    0.03955  -4.031 5.559e-05
```

Le stime per la parte del modello con inflazione di zeri, non danno alcun contributo alla variabile risposta. Per la parte del modello di Poisson valgono gli stessi commenti dell'adattamento precedente. Si calcoli il *TRV* e lo si confronti con un χ^2_2:

```
> 1-pchisq(2*(-58.138270863291 -(-58.4117897249606 )),12-10)
[1] 0.760698
```

Fissato α = 5 %, si accetta senza alcun dubbio l'ipotesi nulla H_0: modello più semplice a sfavore di H_1: modello più complicato. Tuttavia, è consigliato continuare nella selezione dei parametri per la parte del modello contenente alcune covariate che non influiscono sulla variabile risposta. Si elimini, a questo punto, la variabile Gender nella prima parte del modello. Le stime del valore massimo della log-verosimiglianza e dei coefficienti sono:

```
Total Log-likelihood: -58.4407621033055

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1362     0.4293 -0.3171  0.75115
groupnormal   1.5478     0.7736  2.0008  0.04542
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)         -8.3112    2.66268  -3.121 1.800e-03
groupnormal         -1.2868    0.48468  -2.655 7.930e-03
Age                  0.1497    0.03755   3.987 6.685e-05
GenderM             11.3761    2.78550   4.084 4.426e-05
Smokeyes            -0.4718    0.19079  -2.473 1.340e-02
groupnormal:GenderM  1.2911    0.56989   2.266 2.348e-02
Age:GenderM         -0.1584    0.03955  -4.004 6.225e-05
```

A questo punto, solo l'intercetta della parte con inflazione di zeri appare non significativa. Si verifichi quale dei due modelli, il precedente o il corrente, è da ritenersi migliore:

```
> 1-pchisq(2*(-58.4117897249606-(-58.4407621033055)),10-9)
[1] 0.8097743
```

Il modello a cui si è giunti sembra essere il migliore per quanto riguarda la significatività dei parametri. Inoltre, tra tutti i modelli la log-verosimiglianza assume valore massimo in questo modello e il *TRV* è l'ulteriore conferma che il modello scelto è soddisfacente. Si concluda la modellazione dei dati considerando compatibile con i dati la regressione contenente la variabile group per la parte con inflazione di zeri e le variabili `group`, `Age`, `Gender`, `Smoke`, le interazioni tra `group` e `Gender`, tra `Age` e `Gender` per la parte poissoniana del modello. Si può dire, a questo punto, che la probabilità di non avere *spot* di silicio, ossia la probabilità di avere 0 *spot*, dipende dal gruppo di appartenenza; il logaritmo della media del numero di *spot* dipende, invece, dal gruppo di appartenenza, dall'età, dal genere, dall'attitudine al fumo, dalle interazioni tra il gruppo di appartenenza e il genere e tra l'età e il genere. Formalmente, il modello si interpreta mediante le stime dei coefficienti β e γ. Avendo posto $\log(\lambda)=\eta_P$ e $\log it(p)=\eta_B$ si ha:

$$\hat{\eta}_P = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6}$$

$$\hat{\eta}_B = \hat{\gamma}_0 + \hat{\gamma}_1 z_{i1}$$

ossia la formula del modello diventa:

1. $\hat{\eta}_P = \log(\hat{\lambda}) = \log($media `n.spot) = -8.3112 + (-1.2868) groupnormal + 0.1497 Age + 11.3761 GenderM + (-0.4718) Smokeyes + 1.2911 groupnormal:GenderM + (-0.1584) Age:GenderM`.

Ossia la media di `n.spot` si ottiene ricavando esplicitamente il parametro λ secondo la formula (2) riportata al Paragrafo 1.2, da cui risulta $\hat{\lambda} = \exp(\hat{\eta}_P)$.

2. $\hat{\eta}_B = \log it(\hat{p}) =$ logit(probabilità che `n.spot = 0) = -0.1362 + 1.5478 groupnormal`.

La probabilità di avere 0 *spot* si ottiene, usando la formula (2) relativamente a *p*, ossia $\hat{p} = \dfrac{e^{-0.1362 + 1.5478\,\text{groupnormal}}}{1 + e^{-0.1362 + 1.5478\,\text{groupnormal}}}$.

La probabilità di avere 0 *spot* di silicio per un soggetto deceduto per il tumore al polmone è `exp(-0.1362)/(1+exp(-0.1362)) = 0.4660025`, ossia la probabilità di aver *spot* è `(1-0.4660025) = 0.5339975`. Se un soggetto, invece, appartiene al gruppo dei pazienti deceduti per cause naturali, la probabilità di avere 0 *spot* nel polmone è `exp(-0.1362+1.5478)/(1+exp(-0.1362+1.5478)) = 0.8040182`, ossia la probabilità di avere *spot* è `(1-0.8040182) = 0.1959818`. Un soggetto deceduto per cause naturali ha più probabilità di non avere *spot* rispetto a un paziente del gruppo di controllo. Questo fatto è un chiaro segnale di differenza tra i due gruppi: infatti è ragionevole supporre che soggetti deceduti per tumore abbiano quantità più elevate di *spot* di silicio rispetto all'altro gruppo.

3.6.2 Il Modello con Variabile Risposta `n.positive`

Si indaghi, ora, se le covariate che sono risultate influenti per `n.spot`, sono le stesse che possono essere significative anche per `n.positive`, dato che tali variabili risposta sono legate tra di loro. Si inseriscano nella regressione con variabile risposta `n.positive`, tutte le covariate presenti nell'ultima regressione ritenuta valida per spiegare `n.spot`. Si semplifichi eventualmente il modello successivamente tramite un analisi *backward*.
Il valore massimo per la funzione di log-verosimiglianza è `-53.51`.

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1632     0.4350 -0.3752  0.70749
groupnormal   1.5905     0.7782  2.0440  0.04096
------------------------------------------------------------

Count Model (Poisson)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)          -7.5634    3.08709 -2.4500 0.014285
groupnormal          -0.6604    0.52001 -1.2700 0.204101
Age                   0.1315    0.04360  3.0173 0.002550
GenderM               9.4158    3.24975  2.8974 0.003763
Smokeyes             -0.3350    0.24199 -1.3845 0.166204
groupnormal:GenderM   0.3778    0.65252  0.5789 0.562639
Age:GenderM          -0.1289    0.04622 -2.7891 0.005285
```

Per la parte del modello con inflazione di zeri rimane significativa, come per `n.spot`, la variabile `group` al livello fissato $\alpha = 5$ %. Nella parte poissoniana, alcune covariate non sembrano avere un effetto significativo sulla variabile risposta. Si prosegua l'adattamento del modello ai dati togliendo dalla regressione l'interazione `groupnormal:GenderM`. La regressione porta ad avere la log-verosimiglianza che assume il valore `-53.69`. La stima dei coefficienti risulta essere la seguente:

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1569     0.4352 -0.3605  0.71844
groupnormal   1.5922     0.7776  2.0477  0.04059
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.9954    2.87507  -2.433 0.014970
groupnormal  -0.4250    0.30867  -1.377 0.168564
Age           0.1228    0.04017   3.056 0.002244
GenderM       8.7242    2.97274   2.935 0.003338
Smokeyes     -0.2941    0.23125  -1.272 0.203481
Age:GenderM  -0.1180    0.04156  -2.839 0.004521
```

Tralasciando i commenti per la prima parte dell'*output*, i quali risultano identici al passaggio precedente, si osservino le stime della seconda parte: la variabile `Smoke` non è coinvolta nella spiegazione della variabile `n.positive` a livello fissato α = 5 % e 10 %. Si accetta, dunque, il modello più parsimonioso. Si tolga dal modello precedente la variabile `Smoke`. L'*output* di R risulta con le stime:

```
Total Log-likelihood: -54.5374974084299

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1431     0.4349 -0.3291  0.74210
groupnormal   1.5908     0.7765  2.0488  0.04048
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5171    2.76335  -2.358 0.018353
groupnormal  -0.3570    0.29897  -1.194 0.232388
Age           0.1143    0.03834   2.982 0.002864
GenderM       8.1119    2.85827   2.838 0.004539
Age:GenderM  -0.1095    0.03995  -2.741 0.006130
```

Si ponga attenzione alla modellazione poissoniana: l'appartenenza al gruppo sembra non influire sulla variabile risposta a livello fissato α = 5 % e 10 %. Si prosegua l'adattamento ai dati escludendo dalla regressione la variabile `group`. La stima massima della log-verosimiglianza è `-55.34`. I coefficienti stimati sono:

```
 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1325     0.4361 -0.3039  0.76118
```

```
groupnormal   1.6241     0.7736  2.0993  0.03579
---------------------------------------------------------------

Count Model (Poisson)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.7430    2.67596  -2.146 0.031862
Age           0.1025    0.03689   2.777 0.005483
GenderM       7.5827    2.80191   2.706 0.006804
Age:GenderM  -0.1027    0.03917  -2.621 0.008762
```

In questa regressione tutti i coefficienti del modello sono significativi a livello fissato α = 5 %. Si calcola il *TRV*:

```
> 1-pchisq(2*(-54.5374974084299-(-55.3392420550915)),7-6)
[1] 0.2054094
```

e si accetta il modello più semplice. Volendo mantenere il livello fissato α al 5 % l'ultimo modello stimato sembra essere quello che si adatta meglio ai dati. Ciò significa che la probabilità di non avere zone del polmone positive al silicio dipenderebbe dal gruppo di appartenenza, in altre parole vi è una differenza in termini di possibilità di avere zone positive separatamente per le tipologia di decesso. Il logaritmo della media di `n.positive` dipende dall'età, dal genere e dall'interazione tra l'età e il genere. La stima numerica dei parametri β e γ si ottiene descrivendo i predittori lineari:

$$\hat{\eta}_P = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$

$$\hat{\eta}_B = \hat{\gamma}_0 + \hat{\gamma}_1 z_{i1}$$

con

1. $\hat{\eta}_P = \log(\hat{\lambda}) = \log(\text{media } \texttt{n.positive}) = \texttt{-5.7430 + 0.1025 Age + 7.5827 GenderM + (-0.1027) Age:GenderM}$.

Utilizzando la formula (2) si ottiene il valore per la media di `n.positive` che è pari a $\hat{\lambda} = e^{\text{-5.7430 + 0.1025 Age + 7.5827 GenderM + (-0.1027) Age:GenderM}}$.

2. $\hat{\eta}_B = \log it(\hat{p}) = \log it$ (probabilità di avere 0 `n.positive`) = `-0.1325 + 1.6241 groupnormal`.

Analogamente, per la formula (2), si ricava $\hat{p} = \frac{e^{\text{-0.1325 + 1.6241 groupnormal}}}{1 + \text{e}^{\text{-0.1325 + 1.6241 groupnormal}}}$.

Ciò che è interessante è valutare in quale gruppo è più probabile riscontrare `n.positive`. Si calcoli per prima tale probabilità per il gruppo di controllo: `exp(-0.1325)/(1+exp(-0.1325)) = 0.4669234`, ossia la probabilità di avere zone positive al silicio è `1-0.4669234 = 0.5330766`. I soggetti deceduti per cause naturali hanno probabilità di non avere zone positive pari a `exp(-0.1325+1.6241)/(1+exp(-0.1325+1.6241)) = 0.8163183`, ossia la probabilità di avere zone positive è `0.1836817`. I due gruppi hanno, in modo marcato, probabilità diverse di avere zone positive del polmone positive al silicio.

3.7 Previsioni

Poiché le osservazioni relative ai pazienti esposti sono state escluse dall'analisi esplorativa (*cfr*. Paragrafo 3.3) e, naturalmente, anche dall'adattamento del modello, si desidera provare a vedere se i modelli stimati precedentemente, per `n.spot` e `n.positive`, sono in grado di prevedere tali valori. I valori assunti dalle variabili per i soggetti esposti sono visibili al Paragrafo 3.2. Di seguito sono riportati i valori:

```
  group Age Gender Smoke n.positive n.spot
exposed  60      M   yes         31    236
exposed  49      M   yes         24    184
```

Si inizi, con ordine, prevedendo il numero di *spot* per tali pazienti.

3.7.1 La Previsione di `n.spot`

Si inseriscano i valori assunti dalle covariate `group`, `Age`, `Gender` e `Smoke`, considerando il modello definitivo per la variabile risposta `n.spot` (si veda Paragrafo 3.6.2). Essendo, tali soggetti deceduti a causa del tumore, per la variabile `group` si assegna la modalità '`control`'. Si ottiene la seguente previsione per il primo paziente maschio, con età pari a 60 anni e fumatore:

```
> yhat.n.spot
$lower
[1] 2.448456

$upper
[1] 6.766836

$se
[1] 1.135248

$yhat
      [,1]
1 4.251151

$mu
      [,1]
1 7.961179

$phi
       [,1]
1 0.4660148
```

Il numero di *spot* di silicio previsto dal modello per un paziente con le caratteristiche sopra descritte è `4.251151`. Un intervallo di confidenza per il valore predetto, a livello fissato α = 5 %, è `[2.448456; 6.766836]`. La probabilità che nel polmone di

tale paziente non vengano rilevati *spot*, ossia la probabilità di avere 0 *spot* è `0.4660148` (= `phi`), dunque la probabilità che il soggetto abbia *spot* è (`1 - phi`), ossia `0.5339852` (è una probabilità piuttosto alta). La media di `n.spot` che dovrebbe avere tale paziente è `mu = 7.961179`.
La previsione per il secondo paziente maschio, avente 49 anni e fumatore, è:

```
> yhat.n.spot1
$lower
[1] 2.628736

$upper
[1] 7.725398

$se
[1] 1.335258

$yhat
      [,1]
1 4.675104

$mu
      [,1]
1 8.755119

$phi
       [,1]
1 0.4660148
```

La previsione è molto simile alla precedente: il numero di `n.spot` che dovrebbe essere rilevato è `4.675104`. In media il polmone conterrebbe `8.755119` *spot* e la probabilità di avere *spot* è (`1-phi`) = `0.5339852`. L'intervallo di confidenza a livello fissato α = 5 % è `[2.628736;  7.725398]`. Entrambe le previsioni sono molto lontane dal valore vero assunto dai due pazienti esposti (anche l'intervallo di

confidenza non comprende i valori reali). Probabilmente il modello si adatta bene ai dati ma non rispecchia la realtà, poiché purtroppo, la numerosità campionaria a disposizione per la stima dei coefficienti e delle variabili necessarie per spiegare la variabile risposta è piuttosto limitata.

3.7.2 La Previsione di `n.positive`

Analogamente ai valori previsti per `n.spot`, si desidera prevedere i valori che assumerebbe la variabile risposta `n.positive` per gli stessi pazienti. Si inizi dal paziente maschio fumatore sessantenne:

```
> yhat.n.positive
$lower
[1] 1.922375

$upper
[1] 5.042535

$se
[1] 0.7987309

$yhat
      [,1]
1 3.314282

$mu
      [,1]
1 6.217186

$phi
       [,1]
1 0.4669161
```

Per tale paziente il numero di zone del polmone risultate positive al silicio dovrebbe essere `3.314282`, la probabilità di avere zone positive è `(1-phi) = 0.5330839`, il numero medio di `n.positive` è `6.217186`. L'intervallo di confidenza per la previsione a livello fissato $\alpha = 5$ % è `[1.922375; 5.042535]`.

Si preveda il valore di `n.positive` per il secondo soggetto esposto (`Age = 49`, `Gender = 'M'`, `Smoke = 'yes'`):

```
> yhat.n.positive1
$lower
[1] 1.751489

$upper
[1] 5.831453

$se
[1] 1.045113

$yhat
      [,1]
1 3.321861

$mu
      [,1]
1 6.231404

$phi
       [,1]
1 0.4669161
```

Il numero di `n.positive` che dovrebbe avere il soggetto è `3.321861`, la probabilità che possegga zone positive è `(1-phi)= 0.5330839`, il numero medio di

`n.positive` è `6.231404`. L'intervallo di confidenza per la previsione a livello fissato α = 5 % è `[1.751489; 5.831453]`.

Le stime di `n.positive`, come è successo per `n.spot`, non coincidono con i valori reali e nemmeno gli intervalli di confidenza contengono tali valori. Le considerazioni fatte per `n.spot` (si veda Paragrafo 3.7.1) valgono anche per `n.positive`.

3.8 Conclusioni

In questo capitolo si è affrontato il problema di adattare il modello ZIP a un *dataset* contenente dati reali rilevati in ambito medico. Si è concluso che le variabili risposta `n.spot` e `n.positive`, sebbene siano strettamente correlate, non sono spiegate complessivamente dalle stesse variabili (la parte con inflazione di zeri contiene la stessa e unica variabile esplicativa `group` significativa al 5 %). Molto probabilmente ciò è dovuto al campione considerato la cui numerosità è piuttosto bassa (n = `45`). A livello fissato α = 5 % entrambi i modelli sono da ritenere appropriati per i soggetti osservati con le caratteristiche rilevate. In entrambi i modelli la probabilità di possedere `n.spot` e `n.positive` è più alta per il gruppo di controllo (*cfr*. Paragrafi 3.6.1 e 3.6.2), perciò si può asserire che la tipologia del decesso 'spiega' la quantità di zone positive al silicio del polmone e, di conseguenza, anche il numero di *spot* ivi rilevato. Le previsioni sui pazienti esposti sembrano essere lungamente dissimili dai valori assunti nel caso reale; ciò è da ricondursi, quasi certamente, alla numerosità campionaria piuttosto ristretta.

Bibliografia

Bortot, P., Ventura, L., e Salvan, A. (2000). *Inferenza statistica: applicazioni con S-PLUS e R*. Cedam, Padova.

Cameron, A. e Trivedi, P. (1998). *Regression analysis of count data*. Cambridge University Press.

Gentle, J.E (2004). Monte Carlo Methods, In *Encyclopedia of Statistical Sciences* Copyright © 2004 by John Wiley & Sons, Inc. All rights reserved.

Hall, D. B. (2000). Zero-Inflated Poisson and Binomial regression with random effects: A case study. *Biometrics* 56, 1030-1039.

Iacus, S.M. e Masarotto, G. (2003). *Laboratorio di statistica con R*. Mc-Graw-Hill.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1-14.

Pace, L. e Salvan, A. (2001). *Introduzione alla Statistica II - Inferenza, verosimiglianza, modelli*. Cedam.

Winkelmann, R. (2000). *Econometric analysis of count data (fourth ed.).* Berlin: Springer-Verlag.

Xiao-Li Meng (2004). EM Algorithm, In *Encyclopedia of Statistical Sciences* Copyright © 2004 by John Wiley & Sons, Inc. All rights reserved.