UNIVERSITÀ DEGLI STUDI DI TRIESTE

FACOLTÀ DI ECONOMIA

CORSO DI LAUREA IN SCIENZE STATISTICHE ED ATTUARIALI

TESI DI LAUREA IN TECNICA ATTUARIALE DELLE ASSICURAZIONI CONTRO I DANNI

MODELLI CON ETEROGENEITÀ PER LA TARIFFAZIONE IN BASE ALL'ESPERIENZA

LAUREANDO:

STEFANO CALCINA

RELATORE:

PROF. **PATRIZIA GIGANTE**

CORRELATORE:

PROF. **LUCIANO SIGALOTTI**

ANNO ACCADEMICO 2000-2001

INTRODUZIONE

Nell'ambito delle assicurazioni dei rami danni, la tariffazione è il procedimento che conduce a determinare i premi da richiedere agli assicurati attraverso la valutazione probabilistica di opportuni numeri aleatori che descrivono la rischiosità di ciascun individuo: il numero dei sinistri, il risarcimento per sinistro e il risarcimento globale.

Ciò si realizza utilizzando le informazioni di cui si dispone riguardo ai diversi assicurati e che si ritengono influenti sulla rischiosità. Tali informazioni sono tipicamente rappresentate da caratteristiche rilevabili a priori e dall'eventuale storia di sinistrosità osservata.

Usualmente, in una prima fase della tariffazione, gli assicurati vengono ripartiti in classi tariffarie sulla base di caratteristiche osservabili a priori che vengono dette *fattori di rischio* o *variabili tariffarie*. In particolare, nel caso dell'assicurazione sulla responsabilità civile autoveicoli (RCA) si possono considerare variabili tariffarie che descrivono caratteristiche del veicolo (potenza, marca, tipo di alimentazione, data di immatricolazione, ecc.), caratteristiche del contraente (età, sesso, anzianità di patente, professione, stato civile, numero di componenti familiari che hanno la patente, ecc.), caratteristiche del contratto (tipo di copertura assicurativa, frazionamento nel pagamento del premio, livello dei massimali, livello della franchigia, ecc.) ed infine altre variabili che rappresentano caratteristiche legate all'ambiente di circolazione e di residenza (provincia di residenza, zona di circolazione, disponibilità di un box, ecc.).

Le variabili tariffarie sono ripartite in *livelli* o *modalità*. Ciascun livello può essere relativo ad una determinazione della variabile - ad esempio un dato valore per l'occupazione del contraente, un dato valore per la sua età - oppure ad una classe di determinazioni. Nel secondo caso, la partizione in classi di determinazioni della variabile tariffaria è usualmente ottenuta attraverso successive dicotomie - ad esempio un dato valore per l'occupazione contro gli

altri valori, oppure età minore di una data soglia prefissata contro età maggiore. Ci si riferisce a quest'ultimo caso con la locuzione "segmentation approach". Il primo approccio è detto "score". I due metodi, score e segmentation, possono essere combinati.

Naturalmente, si può pensare di considerare molte variabili tariffarie, ma bisogna anche ricordare che non tutte le variabili rilevabili a priori sono rilevanti al fine della descrizione del rischio; talune infatti possono riuscire ridondanti quando sono considerate insieme ad altre. Inoltre, occorre tenere presenti le esigenze commerciali ed i problemi di natura tecnica a cui si va incontro utilizzando un numero elevato di informazioni.

Nel procedimento di costruzione di una tariffa, alla fase iniziale in cui vengono individuate alcune variabili osservabili che si ritiene influiscano sulla rischiosità, segue dunque una fase in cui si selezionano tra queste variabili quelle che risultano essere più significative, tenendo conto anche delle esigenze sopra delineate. L'obiettivo è comunque quello di ripartire gli assicurati in classi di rischi omogenei. Si determina, infine, la stima dei valori attesi delle variabili di interesse per le diverse classi tariffarie e quindi la tariffa. Questo è il procedimento per la *tariffazione a priori*.

Naturalmente, all'interno delle collettività omogenee rispetto alle variabili rilevabili a priori, c'è comunque eterogeneità tra i rischi, fenomeno questo dovuto alle caratteristiche ed ai comportamenti individuali. Ad esempio, nel caso delle assicurazioni RCA, fattori che influenzano il numero di sinistri provocati da un assicurato sono l'abilità nella guida, l'atteggiamento più o meno prudente dell'individuo, la conoscenza delle norme del codice della strada, le distanze percorse, le abitudini riguardo l'assunzione di alcolici. Molti studi, si veda ad esempio Lemaire (1977), hanno messo in evidenza che, più che le caratteristiche osservabili a priori, sono proprio questi fattori ad avere maggiore influenza sulla sinistrosità. Indicazioni su questi elementi possono essere tratte, a posteriori, attraverso l'osservazione della sinistrosità degli individui.

Alle procedure di tariffazione a priori viene dunque spesso affiancata una personalizzazione a posteriori o *tariffazione in base all'esperienza* nella quale si considera l'esperienza individuale di sinistrosità di ciascun assicurato al fine di differenziare i premi con l'obiettivo di far pagare un premio commisurato all'effettiva sinistrosità.

Si capisce, dunque, la necessità di introdurre modelli che prevedano la possibilità di aggiornare, sulla base dell'esperienza, le distribuzioni di probabilità delle diverse variabili di interesse per i singoli assicurati. Tramite la distribuzione a posteriori e quella a priori si costruiscono i *coefficienti di aggiornamento*, ad esempio i coefficienti bonus-malus, che consentono la revisione dei premi ottenuti con il modello a priori.

I modelli con eterogeneità consentono di far fronte a questi problemi. In pratica, si introduce per ogni assicurato un parametro specifico dell'individuo, che, se noto, consente di determinare le distribuzioni del numero di sinistri, dei costi per sinistro, del risarcimento globale attribuibili all'individuo. Poiché tale parametro, detto *parametro di rischio*, deve tenere conto di caratteristiche non osservabili, può essere considerato aleatorio. Formalmente, dunque, le distribuzioni del modello con eterogeneità sono misture di distribuzioni condizionate.

Una volta stimato, il modello con eterogeneità può venire utilizzato per effettuare previsioni su dati di tipo longitudinale e consente di realizzare la tariffazione in base all'esperienza. In un approccio puramente bayesiano, la previsione è costituita dai valori attesi dei numeri aleatori che descrivono la sinistrosità dell'individuo, rispetto alle distribuzioni a posteriori che prendono in considerazione la storia dell'individuo. In quest'ottica, la storia dell'individuo viene vista come rivelatrice dell'eterogeneità non osservabile.

Tuttavia effettuare la previsione può non esser semplice; spesso infatti nei modelli con eterogeneità la verosimiglianza può avere un'espressione che rende difficile una trattazione analitica della stessa ai fini di ottenere le stime dei parametri del modello e di calcolare i coefficienti di aggiornamento del premio.

Jean Pinquet si è occupato dell'introduzione e della stima di modelli con eterogeneità in alcuni articoli pubblicati tra il 1996 e il 2001. Per la stesura della tesi ci siamo basati su questi lavori, in particolare abbiamo fatto riferimento al lavoro "Experience Rating through Heterogeneous Models" (2001).

Passiamo ora ad una breve descrizione dei contenuti dei diversi capitoli.

Nel primo capitolo sono illustrati i modelli per i numeri aleatori che descrivono la rischiosità degli individui: il numero di sinistri, il risarcimento per sinistro, il risarcimento globale. Per ciascun numero aleatorio vengono riportate le distribuzioni dei modelli base, quelle in cui non si considera la componente di eterogeneità ma si tiene conto delle determinazioni delle variabili tariffarie utilizzate nel procedimento di tariffazione a priori. I modelli prevedono infatti che uno dei parametri sia legato a tali determinazioni tramite una componente di regressione. Per i modelli base vengono ricavate le espressioni delle stime di massima verosimiglianza dei relativi parametri. Dopo aver introdotto la distribuzione misturante viene descritto il modello con eterogeneità. In particolare, per il numero di sinistri viene considerato il modello di Poisson con diverse distribuzioni misturanti, per il modello con sinistri di q tipi diversi la distribuzione Poisson-gamma. Per il risarcimento per sinistro vengono illustrati i modelli gamma-gamma e log-normale-normale. Infine, il capitolo si conclude con la descrizione del modello per il risarcimento globale.

Il secondo capitolo è dedicato alla tariffazione in base all'esperienza. Vengono calcolati i coefficienti di adeguamento, detti anche coefficienti bonus-malus, per i modelli presentati al Capitolo 1. Dopo aver proposto i coefficienti ottenuti attraverso un approccio puramente bayesiano, vista la difficoltà che talvolta si riscontra nello specificare l'intera distribuzione del parametro di rischio, vengono presentati i coefficienti ottenuti tramite l'approccio della credibilità lineare.

Nel terzo capitolo viene trattato il problema della stima dei modelli con eterogeneità. Viene descritto il metodo di stima proposto da Pinquet che si basa su alcune proprietà degli stimatori di massima verosimiglianza dei parametri del

modello base. Il metodo prevede di calcolare le stime di massima verosimiglianza per il modello base e poi le stime dei parametri della misturante tramite i residui. Gli stimatori cui si perviene con questa tecnica sono consistenti.

Una parte del lavoro svolto riguarda lo sviluppo di una applicazione numerica. Sono stati stimati i modelli con eterogeneità presentati nel primo capitolo a partire da un portafoglio di 44885 assicurati RCA osservati per tre periodi di tempo consecutivi. Inoltre, con i due metodi, revisione bayesiana e credibilità lineare, sono stati costruiti i coefficienti bonus-malus da applicare al premio a priori per realizzare la tariffazione in base all'esperienza. Nell'ultimo capitolo vengono presentati i risultati numerici ottenuti dallo sviluppo delle valutazioni in ambiente SAS.

CAPITOLO 1

MODELLI CON ETEROGENEITÀ E COMPONENTI DI REGRESSIONE

1.1 INTRODUZIONE

All'interno di collettività di assicurati, omogenee rispetto alle informazioni disponibili a priori rimane una notevole eterogeneità tra i comportamenti degli individui e quindi tra le rispettive rischiosità. I modelli di distribuzione per i numeri aleatori che descrivono la rischiosità, che illustreremo in questo capitolo, consentono di tenere conto di questo aspetto attraverso l'introduzione di un *parametro di rischio* che, in qualche modo, riassume le caratteristiche ed i comportamenti dei singoli individui. Per la stesura ci siamo basati su Pinquet (2001a).

Nel prima paragrafo presentiamo lo schema generale utilizzato nella descrizione dei modelli con eterogeneità studiati in questa tesi.

Si considera dapprima un *modello base*, che utilizza le informazioni disponibili a priori, in quanto uno dei parametri della distribuzione dipende da una componente di regressione legata ai valori assunti da un insieme di variabili tariffarie. Si introduce quindi il parametro di rischio, che può essere considerato un elemento aleatorio che tiene conto dei fattori non osservabili a priori che sono rilevanti per la valutazione probabilistica dei numeri aleatori che descrivono la rischiosità. Si perviene così al modello con eterogeneità che risulta essere una

distribuzione mistura di distribuzioni condizionate ai diversi valori possibili del parametro di rischio e al vettore delle determinazioni delle variabili tariffarie. Il modello base può essere pensato come “innestato” nel modello con eterogeneità, in quanto si suppone che esista una determinazione del parametro di rischio per cui le distribuzioni dei due modelli coincidano.

Nei successivi paragrafi, vengono descritti i modelli e i modelli con eterogeneità per i numeri aleatori che vengono utilizzati nell’ambito della tariffazione.

In particolare, per il modello di sinistri viene preso in considerazione il modello di Poisson nelle due ipotesi in cui il parametro di rischio dipenda o meno dal tempo. Come distribuzioni misturanti vengono proposte la distribuzione gamma e la distribuzione normale.

Talvolta può essere opportuno analizzare separatamente sinistri di tipo diverso, ad esempio distinguere i danni alle cose dai danni alle persone. Per tenere conto di questo aspetto, nel Paragrafo 1.5 viene presentato un modello per q tipi di rischi diversi. Il modello descritto è il modello di Poisson con distribuzione misturante normale.

Per quanto riguarda i risarcimenti, come modello base sono proposti il modello gamma e il modello log-normale. Nei modelli che illustreremo, l’eterogeneità viene modellizzata con una distribuzione gamma per i costi di tipo gamma e con una distribuzione normale per i costi log-normali.

Il capitolo si conclude, infine, con la descrizione del modello per il premio puro. In particolare, viene introdotto il modello Poisson composto in cui il numero di sinistri è di tipo poissoniano e i risarcimenti subordinatamente al numero di sinistri sono indipendenti e identicamente distribuiti. Per la distribuzione dei risarcimenti si sono considerati ancora la gamma e la log-normale. Per questo modello vengono introdotte due componenti di eterogeneità: una per il numero di sinistri e una per il costo, in generale non stocasticamente indipendenti.

1.2 DESCRIZIONE DEL MODELLO GENERALE

[1.2]

In questo paragrafo presentiamo lo schema generale dei modelli con eterogeneità che verranno considerati nel seguito al fine di descrivere probabilisticamente gli elementi che intervengono nella tariffazione. Tali modelli possono venire utilizzati in un approccio di adeguamento del premio in base all'esperienza individuale dell'assicurato.

Il modello di partenza è relativo a variabili osservabili di tipo longitudinale ed è chiamato *basic model* o *modello base*.

Per *dati di tipo longitudinale* si intendono dati relativi ad un campione di unità individuali, come ad esempio persone, imprese, regioni, che siano state osservate per diversi periodi.

Nel caso specifico, per il problema della tariffazione in ambito assicurativo, si considera in primo luogo un vettore $Y_i = (Y_{i1},...,Y_{iT_i})$, dove Y_{it} (t=1, …, T_i) è un numero aleatorio oppure un vettore o un processo di numeri aleatori che descrivono il rischio i-esimo, nel periodo t. Ad esempio, per un fissato assicurato i e per il periodo t, può essere $Y_{it} = N_{it}$, dove N_{it} indica il numero di sinistri dell'individuo, oppure, se interessa distinguere tra loro diversi tipi di sinistri, $Y_{it} = (N_{it}^{(1)},..., N_{it}^{(q)})$, dove $N_{it}^{(j)}$ indica il numero di sinistri di tipo j, oppure ancora $Y_{it} = (N_{it}, C_{it1}, C_{it2},...)$, dove C_{itj} indica l'importo del risarcimento per il j-esimo sinistro.

La dipendenza della distribuzione di Y_i dalle caratteristiche tariffarie avviene introducendo una componente di regressione per uno dei parametri della distribuzione. Si considera dunque una sequenza $x_i = (x_{i1},...,x_{iT_i})$, dove x_{it} è il vettore delle componenti di regressione costituite dalle determinazioni delle variabili tariffarie per l'individuo i, nel periodo t.

Indicato con q_1 un parametro vettoriale, sia

$$\ell^0(y_i | q_1, x_i) \qquad (1.2.1)$$

la distribuzione di Y_i dati q_1 e x_i. Si noti che se y_i è il valore osservato di Y_i, la (1.2.1) indica anche la verosimiglianza dell'osservazione.

Il modello base non tiene conto dell'eterogeneità e costituisce il riferimento per la tariffazione a priori.

Introduciamo ora un altro modello per tenere conto dell'eterogeneità dei rischi all'interno di ogni classe tariffaria. In questo caso si suppone che ogni individuo sia caratterizzato da un *parametro di rischio* il cui valore può essere diverso da individuo a individuo. Poiché le caratteristiche osservabili sono già state considerate il parametro di rischio può essere visto come un elemento che tiene conto dei fattori residui, non osservabili, che sono rilevanti per la descrizione di Y_i.

In letteratura, i modelli con eterogeneità vengono classificati in due modi: il cosiddetto *fixed effects model*, nel quale l'effetto di eterogeneità è introdotto assegnando un parametro certo a ciascun individuo, e il *random effects model*, nel quale il parametro specifico di ogni individuo è aleatorio. Il modello che presentiamo si inquadra in questo secondo approccio.

Allora, oltre al vettore x_i delle variabili esogene osservabili, si introduce nel modello un parametro aleatorio, unidimensionale o vettoriale, U_i che rappresenta il parametro di rischio dell'assicurato i-esimo e che introduce il cosiddetto *random effect*.

Se u_i è una delle determinazioni possibili di U_i, la distribuzione di $Y_i | U_i = u_i$, dati q_1 e x_i, viene indicata con

$$\ell^*(y_i | q_1, x_i, u_i).$$

Queste distribuzioni condizionate costituiscono il cosiddetto fixed effects model laddove la componente di eterogeneità individuale u_i costituisce il fixed effect. Segnaliamo che la terminologia che viene utilizzata spesso in letteratura è

quella di *distribuzione per l'individuo reale* in relazione alla $\ell^*(y_i|q_1, x_i, u_i)$, *distribuzione per l'individuo generico* in relazione alla $\ell^*(y_i|q_1, x_i, U_i)$.

Si suppone ancora che la distribuzione di U_i sia assegnata e che questa dipenda da un vettore di parametri q_2. La distribuzione di Y_i che, dati $q = (q_1, q_2)$ e x_i, indichiamo con $\ell(y_i|q, x_i)$, è ottenibile tramite la proprietà di disintegrabilità (o proprietà iterativa) della speranza matematica. Si ha:

$$\ell(\mathrm{y_i}|\theta, \mathrm{x_i}) = \mathrm{E}_{\theta_2}[\ell^*(\mathrm{y_i}|\theta_1, \mathrm{x_i}, \mathrm{U_i})], \text{ con } q = (q_1, q_2), \qquad (1.2.2)$$

dove la speranza matematica è calcolata rispetto alla distribuzione di U_i. Il parametro θ è scritto, per convenienza, come una coppia di vettori.

Una distribuzione nel modello con eterogeneità è dunque definita condizionatamente alle variabili esogene osservabili ed è ricavata come una mistura, rispetto alla distribuzione del parametro aleatorio U_i, delle distribuzioni condizionate ai diversi valori di tale parametro.

Gli U_i, al variare di i, sono supposti indipendenti e identicamente distribuiti, (i.i.d.). Si suppone inoltre che i vettori Y_i siano stocasticamente indipendenti al variare di i. Per quanto riguarda il legame probabilistico tra gli enti aleatori Y_{it}, al variare di t, e tra le componenti di Y_{it}, per t fissato, saremo più precisi nelle descrizioni dei diversi casi particolari che presenteremo nei prossimi paragrafi, qui ci limitiamo a dire che, sia nel basic model sia nel fixed effects model, si suppone che "il passato non sia correlato con il futuro". In particolare, si assume che le variabili che rappresentano i numeri di sinistri abbiano incrementi indipendenti.

La prossima ipotesi introduce un collegamento tra il modello base e il modello con eterogeneità.

Si suppone che esista u^0, determinazione possibile di U_i, tale che:

$$\ell^*(y_i|q_1, x_i, u^0) = \ell^0(y_i|q_1, x_i), \text{ per ogni } y_i, x_i, q_1. \qquad (1.2.3)$$

Inoltre, per tutti i modelli considerati nel seguito, la distribuzione di U_i corrispondente a $q_2 = 0$ è una distribuzione che concentra la massa unitaria sul valore u^0 e ciò per ogni i. Per indicare questa situazione introduciamo la notazione: $U_i \equiv u^0$ [1].

Dalle (1.2.2) e (1.2.3), come conseguenza dell'ultima condizione discende che, posto $\tilde{\theta} = (\theta_1, 0)$, si ha:

$$\ell(y_i \mid \tilde{\theta}, x_i) = \ell^0(y_i \mid \theta_1, x_i), \text{ per ogni } y_i, x_i, q_1. \qquad (1.2.4)$$

Infatti, fissati $y_i, x_i, q_1,$ per la (1.2.2) si ha: $\ell(y_i \mid \tilde{q}, x_i) = E_{q_2}[\ell^*(y_i \mid q_1, x_i, U_i)]$, con $q_2 = 0$. Poiché per $q_2 = 0$ si ha $U_i \equiv u^0$, si ricava $E_{q_2}[\ell^*(y_i \mid q_1, x_i, U_i)] = l^*(y_i \mid q_1, x_i, u^0)$ e, per la (1.2.3), $\ell^*(y_i \mid q_1, x_i, u^0) = \ell^0(y_i \mid q_1, x_i)$.

Perciò il modello base è compreso nel modello generale con eterogeneità da esso derivato.

Negli esempi considerati più avanti, le distribuzioni condizionate ad ogni fissato valore u_i del parametro di rischio appartengono alla famiglia cui appartiene la distribuzione del modello base. La componente u_i ha effetto sul vettore dei parametri della distribuzione, perciò si può scrivere:

$$\ell^*(y_i \mid q_1, x_i, u_i) = \ell^0(y_i \mid q_1 + g(u_i), x_i),$$
$$\ell(y_i \mid q, x_i) = E_{q_2}[\ell^0(y_i \mid q_1 + g(U_i), x_i)],$$

con una scelta opportuna della funzione g. Dunque le distribuzioni nel modello con eterogeneità sono misture di distribuzioni della famiglia di quella del modello base.

Per quanto riguarda gli insiemi dei valori ammissibili per i parametri $q_1, q_2,$ poniamo

---

[1] Più in generale, per indicare l'uguaglianza in distribuzione tra due numeri aleatori nel seguito useremo il simbolo $=_d$.

[1.2]

$$q_1 \in \Theta_1 \subset \Re^{k_1},\ q_2 \in \Theta_2 \subset \Re^{k_2},\ q \in \Theta = \Theta_1 \times \Theta_2 \in \Re^{k}.$$

Se viene utilizzato un approccio semi-parametrico, il vettore q_2 fornisce vincoli sui momenti della distribuzione misturante.

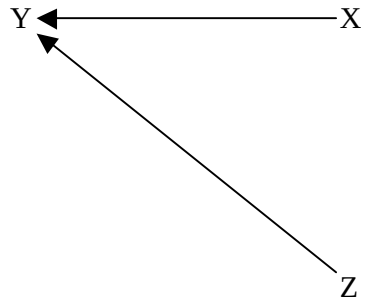
Le distribuzioni misturanti sono solitamente parametrizzate dalle varianze delle componenti del parametro di rischio e dalle covarianze tra le medesime componenti. Lo spazio parametrico Θ_2 è generalmente un cono e $q_2 = 0$ appartiene alla sua frontiera. Ricordiamo che $\Theta_2 \subset \Re^{k_2}$ è un *cono* se da $q_2 \in \Theta_2$ e $l > 0$ segue che $l q_2 \in \Theta_2$.

Concludiamo questo paragrafo sottolineando un aspetto del modello che è stato introdotto: la distribuzione del parametro aleatorio è supposta indipendente dalle determinazioni delle variabili tariffarie. Questo può sembrare in contrasto con le situazioni realistiche nelle quali, in molti casi, si nota una dipendenza tra variabili non osservabili e variabili osservabili. Per esempio, l'età del veicolo (caratteristica osservabile a priori) può essere considerata un buon indicatore della percorrenza annua (caratteristica non osservabile a priori). Ciò si giustifica (si veda Pinquet (2001)) osservando che il prezzo di un veicolo di seconda mano dipende più dalla sua età che dai chilometri percorsi. Quindi si può pensare che una persona che guida poco sia più incentivata a comperare un'automobile di seconda mano e a detenerla il più a lungo possibile. Questo può spiegare anche la significativa influenza dell'età del veicolo sulla frequenza sinistri.

Il contrasto derivante dal fatto che il random effect è qui considerato indipendente dalle variabili esogene osservabili viene superato se il parametro U_i viene visto come elemento che tiene conto dell'eterogeneità residua. Per illustrare questo punto, consideriamo ad esempio un particolare assicurato e un determinato periodo e indichiamo con Y il vettore dei numeri aleatori d'interesse per la descrizione del rischio assicurato. Indichiamo poi con X e Z i vettori aleatori delle variabili tariffarie e delle variabili non osservabili che si ritiene

abbiano rilevanza per la valutazione probabilistica di Y. Supponiamo inoltre che la distribuzione di Y sia determinata se sono note le determinazioni assunte da X e Z.

La relazione di dipendenza tra le variabili è rappresentata nel seguente grafico:



Y: variabili aleatorie di interesse per la descrizione del rischio (endogene)
X: fattori tariffari (esogeni, osservabili)
Z: variabili non osservabili (esogene)

Per fissare le idee, Y rappresenti il numero aleatorio di sinistri dell'assicurato e per la distribuzione di $Y|(X = x, Z = z)$ si ponga:

$$\tilde{\ell}(y|x,z) = P_{\lambda}(y) = \exp(-\lambda)\frac{\lambda^{y}}{y!},\ l = \exp(x'a + z'b), \qquad (1.2.5)$$

dove x', z' indicano i vettori riga trasposti di x, z e α, β sono vettori di parametri[2].

Nell'ambito in cui ci siamo posti, la distribuzione per il modello con eterogeneità può essere espressa tramite la

$$\ell(y|x) = E[\tilde{\ell}(y|X,Z)|X = x] \qquad (1.2.6)$$

dove, dato y, $\tilde{\ell}(y|X,Z)$ è il numero aleatorio funzione della coppia (X,Z) che all'evento ($X = x$, $Z = z$) associa il valore in y della funzione di probabilità o di

[2] Segnaliamo che in questa tesi i vettori sono sistematicamente intesi come vettori colonna. I trasposti di un vettore colonna o di una matrice sono indicati con un apice.

densità di $Y|(X = x, Z = z)$. A partire dal numero aleatorio $\tilde{\ell}(y|X, Z)$ si considera il numero condizionato $\tilde{\ell}(y|X, Z)|X = x$ e di questo si calcola la speranza matematica. Nell'ultimo passaggio si utilizza la distribuzione di $Z|X = x$ che è ottenibile dalla distribuzione congiunta di (X,Z), supposta assegnata.

Ricaviamo ora la regressione lineare di Z rispetto a X, cioè determiniamo $a \in \Re$ tale che

$$Z = aX + V \text{ [3]},$$

con V tale che:

$$E(X'V) = 0 .$$

In altre parole, determiniamo $a \in \Re$, tale che posto $V = Z - aX$, riesca $E(X'V) = 0$.

Poiché $X'V = X'Z - aX'X$, si ricava:

$$a = [E(X'X)]^{-1} E(X'Z) .$$

Possiamo sempre supporre $E(V) = 0$, eventualmente ridefinendo in modo opportuno i vettori Z e X; basta per questo porre $\tilde{Z} = Z - E(Z)$ e $\tilde{X} = X - E(X)$.

Se $E(V) = 0$ riesce:

$$Cov(X, V) = E(X'V) - E(X)E(V) = 0 .$$

Se in particolare la distribuzione congiunta di (X,Z) è tale che X e $V = Z - aX$ siano stocasticamente indipendenti, allora la distribuzione di $Z - aX|(X = x) = V|(X = x)$ è quella di V per ogni x. Pertanto nella (1.2.5), si ha

$$\tilde{\ell}(y|X, Z)|(X = x) = \tilde{\ell}(y|X, aX + V)|(X = x) = \tilde{\ell}(y|x, ax + V|X = x)$$

---

[3] Si osservi che qui si richiede implicitamente che Z e X abbiano uno stesso numero di componenti; possiamo sempre realizzare questa condizione introducendo eventuali componenti certe per uno dei due vettori.

e, poiché la distribuzione di $V|X = x$ è quella di V,

$$\ell(y|x) = E[\tilde{\ell}(y|X, Z)|X = x] = E[\tilde{\ell}(y|x, ax + V)].$$

Definendo opportunamente ℓ^*, θ_1 e U possiamo porre

$$\mathrm{E}[\tilde{\ell}(\mathrm{y}|\mathrm{x}, \mathrm{ax} + \mathrm{V})] = \mathrm{E}[\ell^*(\mathrm{y}|\theta_1, \mathrm{x}, \mathrm{U})],$$

dove la distribuzione di U non dipende da x.

Ad esempio con riferimento alla (1.2.5), si ha $\mathrm{U} = \mathrm{V}'\beta$ e $\ell^*(y|q_1, x, u) = P_l(y)$, con $l = \exp(x'q_1 + u)$ e $q_1 = a + ab$.

Infatti, in questo caso il numero aleatorio $\tilde{\ell}(y|x, ax + V)$ associa a $V = v$ il valore in y della distribuzione di Poisson di parametro $l = \exp[x'a + (ax + v)'b] = \exp[x'(a + ab) + v'b]$. Dunque se poniamo $q_1 = a + ab$ e $u = v'b$, il parametro della distribuzione può essere scritto come sopra indicato.

Poiché $U = V'b$ è indipendente da X, il random effect riceve l'interpretazione data in precedenza.

Nei prossimi paragrafi presentiamo alcuni esempi di modelli con eterogeneità che possono essere usati per effettuare la tariffazione sulla base dell'esperienza e che rispecchiano lo schema generale che abbiamo illustrato. Questi modelli sono ricavati a partire da modelli base che diremo anche *modelli a priori* per i quali vengono richiamate alcune proprietà fondamentali.

1.3 MODELLI PER IL NUMERO DI SINISTRI

Con riferimento all'individuo i, $i = 1, \ldots, p$, indichiamo con N_{it} il numero dei sinistri riportati nel periodo t.

Consideriamo come modello base il modello di Poisson:

$$N_{it} \sim P(l_{it}), \text{ con } l_{it} = \exp(x'_{it}q_1) \quad i = 1,...,p; \qquad t = 1,2,....$$

Il parametro della distribuzione dipende dunque dalle caratteristiche tariffarie nel periodo t, che sono supposte note, tramite la componente di regressione $x'_{it}q_1$.

Ricordiamo che, come indicato nello schema generale, i processi $(N_{i1}, N_{i2},...)$ sono stocasticamente indipendenti, al variare di i, e, per ogni fissato i, i numeri aleatori del processo corrispondente sono stocasticamente indipendenti.

Con i simboli introdotti nel Paragrafo 1.2 si ha che $Y_i = (N_{i1},...,N_{iT_i})$ e

$$P(N_{it} = n|q_1, x_i) = e^{-l_{it}} \frac{l_{it}^n}{n!}, \; l_{it} = \exp(x'_{it}q_1),$$

è la distribuzione marginale di N_{it} relativa alla $\ell^0(y_i|q_1, x_i)$.

Supponendo di disporre delle osservazioni dei numeri di sinistri riportati dai p rischi in più periodi e indicato con n_{it} il numero di sinistri denunciati dal rischio i nel periodo t, la stima di massima verosimiglianza di q_1 risolvendo l'equazione vettoriale:

$$\sum_{i,t}(n_{it} - l_{it})x_{it} = 0, \tag{1.3.1}$$

dove $l_{it} = \exp(x'_{it}q_1)$.

Infatti, la verosimiglianza è data da

$$\ell = \prod_{i,t} P(N_{it} = n_{it}|q_1, x_i) = \prod_{i,t} e^{-l_{it}} \frac{l_{it}^{n_{it}}}{n_{it}!}.$$

Passando alla log-verosimiglianza, si ha

$$l = \sum_{i,t}\left(-l_{it} + n_{it}\log l_{it} - \log(n_{it}!)\right).$$

Calcolando la derivata di l rispetto a q_1, ovvero il vettore delle derivate parziali di l rispetto alle componenti di θ_1, si verifica facilmente che riesce

$$\frac{\partial l}{\partial q_1} = \sum_{i,t} \left(- l_{it} x_{it} + n_{it} x_{it} \right) = \sum_{i,t} \left(n_{it} - l_{it} \right) x_{it} .$$

Ricordiamo che l_{it} è funzione di q_1, e quindi anche l è funzione di q_1.

Infine, dalla condizione $\frac{\partial l}{\partial q_1} = 0$, si ricava la (1.3.1). Ricordiamo che il vettore $\frac{\partial l}{\partial q_1}$ è detto *score*.

Ottenuta la stima di q_1, $\hat{q}_1$, la stima del valore atteso di N_{it} è $\hat{l}_{it} = \exp(x'_{it}\hat{q}_1)$. Osserviamo che, se per le polizze si hanno periodi di osservazioni di durate diverse, occorre considerare le esposizioni. Per il momento però non teniamo conto di questo aspetto.

Si noti che la (1.3.1) può essere vista come una relazione di ortogonalità tra le componenti di regressione e i residui: $n_{it} - l_{it}$.

Se le variabili tariffarie sono dicotomiche, i vettori x_{it} sono sequenze costituite dalle cifre 0 e 1. In questo caso la (1.3.1) richiede che, considerando il sottoinsieme delle osservazioni corrispondenti ad un fissato livello di una variabile tariffaria, la somma dei numeri di sinistri osservati sia uguale alla somma delle frequenze stimate. Infatti, indicata con $x_{it}^{(j)}$ la j-esima componente del vettore x_{it}, la condizione (1.3.1) è equivalente alle

$$\sum_{i,t} (n_{it} - l_{it}) x_{it}^{(j)} = 0, \text{ per ogni j.}$$

Ora, si ha

$$\sum_{i,t} (n_{it} - l_{it}) x_{it}^{(j)} = \sum\nolimits_{i,t}^{(1)} (n_{it} - l_{it}) = \sum\nolimits_{i,t}^{(1)} n_{it} - \sum\nolimits_{i,t}^{(1)} l_{it} ,$$

dove $\sum_{i,t}^{(1)}$ indica la somma estesa a tutti gli addendi per cui $x_{it}^{(j)} = 1$.

La (1.3.1) comporta dunque una condizione di *bilanciamento* del modello tariffario.

Introduciamo ora il modello con eterogeneità.

Distinguiamo due casi.

1) LA COMPONENTE DI ETEROGENEITÀ NON DIPENDE DAL TEMPO

Nell'approccio del fixed effects model, è fissato u_i e si suppone che

$$N_{it} \sim P(l_{it} w_i) \qquad t = 1,...,T_i,\ i = 1,...,p,$$

dove $w_i = g(u_i)$, con g funzione opportuna, e $l_{it} = \exp(x_{it}' q_1)$.

Nell'approccio del random effects model, si assume $W_i = g(U_i)$ e

- $N_{it} | U_i = u_i \sim P(l_{it} w_i)$, $w_i = g(u_i)$, $l_{it} = \exp(x_{it}' q_1)$,
- $N_{it} | U_i = u_i$ stocasticamente indipendenti.

Inoltre, al variare di i, i processi $(U_i, N_{i1}, N_{i2}, ...)$ sono stocasticamente indipendenti.

Osserviamo che, con i simboli del modello generale, $P(N_{it} = n | q_1, x_i, u_i) = e^{-l_{it} w_i} \frac{(l_{it} w_i)^n}{n!}$ è la distribuzione marginale di $N_{it} | U_i = u_i$ relativa alla $\ell^*(y_i | \theta_1, x_i, u_i)$.

La funzione g è generalmente la funzione identica. Tuttavia, per alcune scelte della distribuzione di U_i è opportuno pensare a trasformate di U_i, ad esempio $W_i = g(U_i) = e^{U_i}$, si veda il prossimo Esempio 2.

La distribuzione del parametro aleatorio viene scelta in opportune famiglie parametriche. In particolare, il parametro è rappresentato dalla varianza.

Scegliendo come distribuzione del parametro aleatorio la coniugata naturale della verosimiglianza, si può ricavare in modo analitico la distribuzione di N_{it}.

- *Esempio 1.*

La distribuzione Gamma è coniugata della Poisson. Si abbia $W_i = U_i \sim \text{gamma}(\delta,\delta)$, dunque i numeri aleatori U_i hanno speranza matematica unitaria e varianza pari a $\sigma^2 = \frac{1}{\delta}$. Poniamo $U_i \sim \gamma(\sigma^2)$, quindi $\theta_2 = \sigma^2$.

Poiché $N_{it} | U_i = u_i \sim P(l_{it} u_i)$, in questo caso, nella (1.2.3), si ha $u^0 = 1$. Le distribuzioni dei numeri aleatori N_{it}, nel modello con eterogeneità, sono binomiali negative. Questo è un modello molto utilizzato nella letteratura attuariale poiché la verosimiglianza è analiticamente trattabile, ed inoltre i coefficienti bonus-malus sono espliciti e facilmente interpretabili con riferimento alla credibilità attribuita alla storia dell'assicurato. Infatti, posto $H = (N_{i1} = n_{i1},...,N_{it} = n_{it})$ e indicate con f_{U_i}, $f_{U_i|H}$ le densità di U_i e $U_i|H$ si ha

$$f_{U_i|H}(u) \propto f_{U_i}(u) \prod_{h=1}^{t} e^{-l_{ih}u} \frac{(l_{ih}u)^{n_{ih}}}{n_{ih}!} \propto$$

$$\propto u^{d-1} e^{-ud} e^{-u\sum_{h=1}^{t} l_{ih}} u^{\sum_{h=1}^{t} n_{ih}} = u^{(d+\sum_{h=1}^{t} n_{ih})-1} e^{-u(d+\sum_{h=1}^{t} l_{ih})}.$$

Quindi si ricava che la distribuzione a posteriori di U_i, visto H, è la $gamma(d + \sum_{h=1}^{t} n_{ih}, d + \sum_{h=1}^{t} l_{ih})$. Per la proprietà di disintegrabilità, la speranza matematica di $N_{it+1}|H$ è

$$E[N_{it+1}|H] = \int_0^{+\infty} E[N_{it+1}|H, U_i = u] f_{U_i|H}(u) du = l_{it+1} \int_0^{+\infty} u f_{U_i|H}(u) du =$$

$$= l_{it+1} E(U_i|H).$$

Si ha quindi

$$E[N_{it+1}|N_{i1}=n_{i1},...,N_{it}=n_{it}]=l_{it+1}\frac{d+\sum_{h=1}^{t}n_{ih}}{d+\sum_{h=1}^{t}l_{ih}}. \qquad (1.3.2)$$

Osservando che $E(N_{it+1})=l_{it+1}$, per il cosiddetto *coefficiente bonus-malus* o *coefficiente di adeguamento in base all'esperienza* si ha

$$\frac{E[N_{it+1}|N_{i1}=n_{i1},...,N_{it}=n_{it}]}{E[N_{it+1}]}=\frac{l_{it+1}E[U_{it+1}|N_{i1}=n_{i1},...,N_{it}=n_{it}]}{l_{it+1}}=\frac{d+\sum_{h=1}^{t}n_{ih}}{d+\sum_{h=1}^{t}l_{ih}}.$$

Questo coefficiente è interpretabile in termini di credibilità attribuita alla storia dell'assicurato poiché riesce

$$\frac{E[N_{it+1}|N_{i1}=n_{i1},...,N_{it}=n_{it}]}{E[N_{it+1}]}=(1-a)\cdot 1+a\frac{\sum_{h=1}^{t}n_{ih}}{\sum_{h=1}^{t}l_{ih}}$$

con $a=\frac{\sum_{h=1}^{t}l_{ih}}{\sum_{h=1}^{t}l_{ih}+d}$. Il coefficiente bonus-malus in t+1 è dunque media ponderata del valore a priori, pari a 1, e del rapporto tra il numero di sinistri osservato nei periodi da 1 a t e il numero atteso, valutato a priori, di tali sinistri.

- *Esempio 2.*

Consideriamo qui il caso in cui si assuma per il parametro di rischio U_i una distribuzione normale.

La normale non è coniugata della Poisson. Tuttavia scegliere la distribuzione normale per U_i e porre $W_i=e^{U_i}$ può riuscire di notevole interesse in relazione ai risultati che si ottengono. Osserviamo che in questo caso W_i ha distribuzione log-normale. In particolare, si pone $U_i \sim N(0,s^2)$. Si ha

[1.3]

$N_{it}|U_i = u_i \sim P(l_{it} w_i)$ con $l_{it} w_i = \exp(x'_{it} q_1 + u_i)$ e perciò $u^0 = 0$ nell'equazione (1.2.3). La verosimiglianza non è analiticamente trattabile ed i coefficienti bonus-malus non sono determinabili in modo esplicito. Il vantaggio però consiste nel fatto che questo modello si presta ad essere generalizzato introducendo la dipendenza dal tempo nelle componenti di eterogeneità. Inoltre la distribuzione gaussiana si estende in modo naturale al caso multivariato, conducendo a modelli con eterogeneità nel caso in cui N_{it}, anziché un numero aleatorio, è un vettore con componenti i numeri di sinistri di diversi tipi. Il primo di questi due aspetti viene messo in evidenza nel prossimo sottoparagrafo, il secondo nel Paragrafo 1.4.

2) LA COMPONENTE DI ETEROGENEITÀ DIPENDE DAL TEMPO

È piuttosto naturale introdurre una generalizzazione di questo tipo dal momento che variabili non osservabili e rilevanti per la descrizione del rischio possono modificarsi con il tempo così come accade per le variabili osservabili.

Si pensi alla variazione nei comportamenti di un individuo che possono derivare da un lato dal verificarsi in determinati periodi di particolari eventi, come per esempio il divorzio oppure il licenziamento, e dall'altro da situazioni che producono i loro effetti nel tempo, come ad esempio la dipendenza dall'assunzione di alcolici. I due modelli che seguono sono adatti a descrivere rispettivamente, le situazioni dei due tipi indicati.

Osserviamo che i due modelli possono essere adottati anche per descrivere gli effetti prodotti da cause "endogene", come ad esempio modificazioni nei comportamenti dovute alla presenza di incentivi prodotti da una tariffazione in base all'esperienza (si veda il Capitolo 2).

Indicato con U_{it} il parametro aleatorio per l'individuo i, nel periodo t, e posto $W_{it} = g(U_{it})$, si richiede che

- $N_{it}|U_{it} = u_{it} \sim P(l_{it} w_{it})$, $w_{it} = g(u_{it})$.

- $N_{it}\left|(U_{i1}=u_{i1},...,U_{iT_i}=u_{iT_i}\right.)=_d N_{it}\left|U_{it}=u_{it}\right.$, $t=1,...,T_i$
- i numeri aleatori del processo $N_{it}\left|\left(U_{i1}=u_{i1},...,U_{iT_i}=u_{iT_i}\right.\right)$, $t=1,...,T_i$, siano stocasticamente indipendenti.

Si assume infine indipendenza stocastica tra i processi relativi ai diversi rischi del portafoglio.

Per quanto riguarda la distribuzione del processo $\{U_{it}, t \geq 0\}$, presentiamo due possibili scelte.

- *Esempio 3.*

Si abbia $U_{it}=W_{it}=R_iS_{it}$. Il parametro aleatorio U_{it} è espresso come prodotto di due numeri aleatori: R_i che dipende solo da i e S_{it} che dipende anche da t. I numeri aleatori R_i sono i.i.d. al variare di i, i numeri aleatori S_{it} sono i.i.d. al variare di i e t e i due processi $\{R_i, i=1,..., p\}$, $\{S_{it}, i=1,..., p; t=1,2,...\}$ sono stocasticamente indipendenti.

In questo modello, la funzione di autocorrelazione tra i parametri aleatori è costante. Infatti si ha

$$Cov(U_{it},U_{it+h})=E(U_{it}U_{it+h})-E(U_{it})E(U_{it+h})=E(R_i^2S_{it}S_{it+h})-E(R_iS_{it})E(R_iS_{it+h})=$$
$$=E(R_i^2)m_1^2-E(R_i)^2m_1^2=m_1^2Var(R_i),$$

dove $m_1=E(S_{it})$ è costante al variare di t per l'ipotesi di uguale distribuzione dei numeri aleatori S_{it}. I passaggi sono conseguenza delle ipotesi di indipendenza stocastica tra i numeri aleatori coinvolti.

Inoltre,

$$Var(U_{it})=E(U_{it}^2)-E(U_{it})^2=E(R_i^2S_{it}^2)-(E(R_i)E(S_{it}))^2=$$
$$=E(R_i^2)m_2-E(R_i)^2m_1^2=Var(U_{it+h}),$$

dove $m_2=E(S_{it}^2)$.

[1.3]

Indicata ora con $r(t,t+h)$ la funzione di autocorrelazione del processo $\{U_{it}, t \geq 0\}$ ovvero

$$r(t,t+h) = \frac{Cov(U_{it}, U_{it+h})}{\sqrt{Var(U_{it})Var(U_{it+h})}},$$

si ottiene

$$r(t,t+h) = \frac{m_1^2 Var(R_i)}{E(R_i^2)m_2 - E(R_i)^2 m_1^2}$$

che risulta essere appunto costante: non dipende né da t né da t+h.

Il processo che introduce la dipendenza dal tempo, $(S_{it})_{t=1,\ldots,T_i}$, consente di modellizzare l'effetto di eventi fortuiti ("shock") nella distribuzione dell'individuo. La specificazione che viene utilizzata in questo caso per le due componenti aleatorie R_i e S_{it} è semi-parametrica. Il modello viene ripreso nel Paragrafo 8.1.

- *Esempio 4.*

Poniamo $W_{it} = e^{U_{it}}$ dove, per ogni i, la distribuzione di $\{U_{it}, t \geq 0\}$ è quella di un processo stazionario gaussiano. In altre parole, oltre alla stazionarietà si suppone che $U_{it} \sim N(0, s^2)$. In questo caso la funzione di autocorrelazione non dipende dai particolari istanti considerati, ma solo dalla loro "distanza". Si ha

$$r(t,t+h) = r(h) = \frac{\text{cov}(U_{it}, U_{it+h})}{s^2},$$

che dipende da h, ma non da t, per la stazionarietà del processo. Questa proprietà della funzione di autocorrelazione fa sì che il modello sia adatto a descrivere situazioni nelle quali la frequenza sinistri dipende dal tempo trascorso rispetto ad un fissato periodo.

Osservazione. *Il modello di Hausman.*

Il modello descritto nell'Esempio 3 segue lo stesso approccio del modello binomiale negativo con random effects proposto da Hausman et al. (1984), che descriviamo in questa osservazione.

Il modello si ottiene a partire dalle distribuzioni condizionate definite come segue:

$$N_{it}\,|\Lambda_{it} = l_{it} \sim \mathrm{P}(\lambda_{\mathrm{it}}) \text{ con } \Lambda_{it} \sim gamma(m_{it}, d_i) \text{ e } m_{it} = \exp(x'_{it} b),$$

dove $gamma(m_{it}, d_i)$ indica la distribuzione gamma di speranza matematica $\frac{m_{it}}{d_i}$ e varianza $\frac{m_{it}}{d_i^2}$. Dunque m_{it} è il parametro di forma della distribuzione. Si osservi che le componenti di regressione sono incluse nel parametro di forma.

Poniamo $G_{it} = d_i \Lambda_{it}$. Allora, $G_{it} \sim gamma(m_{it}, 1)$ e si può scrivere:

$$\Lambda_{it} = \frac{G_{it}}{d_i}.$$

Riesce allora $E(N_{it}) = E(\Lambda_{it}) = \frac{m_{it}}{d_i}$ e N_{it} ha distribuzione binomiale negativa di parametri m_{it}, $\frac{d_i}{d_i + 1}$.

Nel modello binomiale negativo con random effects, d_i è visto come la realizzazione di un numero aleatorio Δ_i con

$$\Delta_i = \frac{A_i}{B_i} \text{ dove } A_i \sim gamma(a,1),\ B_i \sim gamma(b,1).$$

Inoltre i numeri aleatori del processo $(A_i, B_i, G_{it})_{\substack{i=1,\ldots p \\ t=1,\ldots,T_i}}$ sono supposti indipendenti.

In modo più preciso, possiamo descrivere il modello come segue. Per l'individuo i, fissato, si assume che la distribuzione del processo $(N_{it})_{t=1,\ldots,T_i}$

dipenda dal processo aleatorio $(A_i, B_i, G_{it}; t = 1,...,T_i)$, e che posto $\Delta_i = \frac{A_i}{B_i}$, $\Lambda^*_{it} = \frac{G_{it}}{\Delta_i}$, $G_i = (G_{i1},..,G_{iT_i})$, $g_i = (g_{i1},...,g_{iT_i})$, si abbia

$$N_{it}|(\Delta_i = d_i, G_i = g_i) =_d N_{it}|(\Delta_i = d_i, G_{it} = g_{it}) =_d N_{it}\left|\left(\Delta_i = d_i, \Lambda^*_{it} = \frac{g_{it}}{d_i}\right) \sim P\left(\frac{g_{it}}{d_i}\right).\right.$$

Inoltre, al variare di t, i numeri aleatori $N_{it}|(\Delta_i = \delta_i, G_i = g_i)$ siano stocasticamente indipendenti e per $(A_i, B_i, G_{it}; t = 1,...,T_i)$ sussistano le ipotesi precedentemente indicate. Allora, il processo $\left(N_{it}|\Delta_i = d_i\right)_{i=1,...,T_i}$ è mistura dei processi $\left(N_{it}|\Delta_i = d_i, G_i = g_i\right)_{t=1,...,T_i}$.

In particolare, la distribuzione di $N_{it}|\Delta_i = d_i$ è binomiale negativa in quanto mistura di poissoniana con misturante la distribuzione di $\Lambda^*_{it}|\Delta_i = \delta_i$ che è la $gamma(m_{it}, d_i)$.

Si assume poi, l'indipendenza tra i processi relativi ai diversi individui del portafoglio.

Questo modello può esser visto come un modello di Poisson con componenti di eterogeneità dipendente dal tempo, del tipo $U_{it} = R_i S_{it}$. Infatti, Λ^*_{it} può essere scritto nel seguente modo:

$$\Lambda^*_{it} = E(\Lambda^*_{it}) R_i S_{it}, \text{ con } R_i = \frac{1/\Delta_i}{E(1/\Delta_i)} \text{ e } S_{it} = \frac{G_{it}}{E(G_{it})}.$$

Per provare questa affermazione osserviamo che si ha

$$E\left(\frac{1}{\Delta_i}\right) = E\left(\frac{B_i}{A_i}\right) = \frac{b}{a-1},$$

dove l'ultima uguaglianza segue dalla

$$E(A_i^{-1}) = \frac{1}{\Gamma(a)} \int_0^{+\infty} x^{a-2} e^{-x} dx = \frac{\Gamma(a-1)}{\Gamma(a)} = \frac{1}{a-1}$$

e dall'indipendenza stocastica tra A_i e B_i. Allora

$$E(\Lambda_{it}^*) = E\left(\frac{G_{it}}{\Delta_i}\right) = E(G_{it})E\left(\frac{1}{\Delta_i}\right) = m_{it}\frac{b}{a-1}.$$

Quindi possiamo porre

$$\Lambda_{it}^* = \frac{G_{it}}{\Delta_i} E(\Lambda_{it}^*)\frac{a-1}{bm_{it}} = E(\Lambda_{it}^*)\frac{1/\Delta_i}{E(1/\Delta_i)}\frac{G_{it}}{E(G_{it})}.$$

Inoltre, si ha

$$E(R_i) = E(S_{it}) = 1 \text{ per ogni } i,t.$$

Ricaviamo ora la varianza di R_i e S_{it}. Dalle

$$E(B_i^2) = \frac{1}{\Gamma(b)} \int_0^{+\infty} x^{b+1} e^{-x} dx = \frac{\Gamma(b+2)}{\Gamma(b)} = (b+1)b,$$

$$E(A_i^{-2}) = \frac{1}{\Gamma(a)} \int_0^{+\infty} x^{a-3} e^{-x} dx = \frac{\Gamma(a-2)}{\Gamma(a)} = \frac{1}{(a-1)(a-2)}$$

si ricava:

$$E\left[\left(\frac{1}{\Delta_i}\right)^2\right] = E\left(\frac{B_i^2}{A_i^2}\right) = \frac{E(B_i^2)}{E(A_i^2)} = \frac{(b+1)b}{(a-1)(a-2)}.$$

Con semplici passaggi si ottiene:

$$Var(R_i) = \frac{a+b-1}{b(a-2)}.$$

Per S_{it} si ha

[1.3]

$$Var(S_{it}) = Var\left(\frac{G_i}{E(G_i)}\right) = \frac{1}{m_{it}} = \frac{b}{(a-1)E(N_{it})}.$$

Per ottenere l'ultima uguaglianza, ricordiamo che $N_{it}|\Delta_i = d_i$ ha distribuzione binomiale negativa di parametri $\left(m_{it}, \frac{d_i}{d_i+1}\right)$, dunque $E\left(N_{it}|\Delta_i = d_i\right) = \frac{m_{it}}{d_i}$. Allora si ha

$$E(N_{it}) = \int E(N_{it}|\Delta_i = d)dF_{\Delta_i}(d) = m_{it}E\left(\frac{1}{\Delta_i}\right) = m_{it}\frac{b}{a-1}$$

da cui si ricava $m_{it} = E(N_{it})\frac{a-1}{b}$.

Osserviamo che la componente non dipendente dal tempo, R_i, ha una varianza finita se $a > 2$. Si può vedere che la verosimiglianza inoltre è trattabile analiticamente (si veda Hausman et al. (1984)).

Ritornando ora al modello generale, ricaviamo l'espressione dei momenti del secondo ordine per i numeri aleatori del processo N_{it} con componente di eterogeneità dipendente dal tempo e $W_{it} = U_{it}$. Riesca inoltre $E(U_i) = 1$.

Ricordiamo, per questo, che le formule per la varianza e la covarianza nei modelli mistura sono

$$Var(N) = E_U[Var(N|U)] + Var_U[E(N|U)],$$

$$Cov(N_1, N_2) = E_U[Cov(N_1, N_2|U)] + Cov_U[E(N_1|U), E(N_2|U)], \qquad (1.3.2)$$

dove E_U, Var_U e Cov_U si riferiscono a speranze matematiche calcolate rispetto alla distribuzione di U.

Per l'attuale modello

$$Var(N_{it}) = E[l_{it}U_{it}] + Var[l_{it}U_{it}] = l_{it}E[U_{it}] + l_{it}^2 Var[U_{it}] = l_{it} + l_{it}^2 Var[U_{it}]$$

[1.3]

quindi

$$Var(N_{it}) - E(N_{it}) = l_{it}^2 Var(U_{it}) .$$

Pertanto, si ha $Var(N_{it}) > E(N_{it})$: i modelli con eterogeneità comportano *sovradispersione*. Prima di adottare un tale modello occorre quindi verificare se i dati disponibili supportano l'ipotesi di sovradispersione e sono sufficienti per realizzare la stima dei parametri. Un'analisi del problema nel caso in cui la componente di eterogeneità sia costante e la distribuzione mistura sia parametrizzata dalla varianza è riportata nel Paragrafo 3.2.

Per quanto riguarda le covarianze si ha

$$Cov(N_{it}, N_{it'}) = E[Cov(N_{it}, N_{it'} | U_{it}, U_{it'})] + Cov[E(N_{it} | U_{it}, U_{it'}), E(N_{it'} | U_{it}, U_{it'})] =$$
$$= Cov[E(N_{it} | U_{it}), E(N_{it'} | U_{it'})] = l_{it} l_{it'} Cov(U_{it}, U_{it'})$$

I modelli con eterogeneità introducono, dunque, correlazione tra i numeri del processo $\{N_{it}, t \geq 0\}$.

1.4 MODELLI PER IL NUMERO DI SINISTRI DI DIVERSI TIPI

A volte, in presenza di sinistri di diversi tipi, può essere interessante analizzare un rischio anche in relazione al tipo di sinistri riportati. A tale proposito, si pensi ad esempio nelle assicurazioni RCA alla distinzione tra danni a cose e lesioni a persone o anche tra i due precedenti tipi di danni e altri eventi, come infrazioni al codice della strada.

Supponiamo dunque di disporre, per ogni individuo, di un processo di osservazione vettoriale relativo ai sinistri di q tipi di rischi diversi.

Sia $N_{it}^{(j)}$ il numero aleatorio di sinistri di tipo *j* denunciati nel periodo *t* dall'individuo *i* e $N_{it} = \left(N_{it}^{(1)}, ... N_{it}^{(q)}\right)$. Preso il modello di Poisson come modello base, si ha

[1.4]

$$N_{it}^{(j)} \sim P\left(l_{it}^{(j)}\right), \text{ con } \; l_{it}^{(j)} = \exp\left(x_{it}'^{(j)} q_{1j}\right), \; i = 1,\ldots, p; \; t = 1,2,\ldots; \; j = 1,\ldots, q .$$

Si suppone che i numeri aleatori $N_{it}^{(j)}$ siano stocasticamente indipendenti al variare di *j*. Inoltre si suppone che, fissato *j*, i numeri aleatori del processo $\left(N_{i1}^{(j)}, N_{i2}^{(j)},\ldots\right)$ siano stocasticamente indipendenti come indicato nello schema generale, supponiamo poi l'indipendenza tra i processi relativi ai diversi individui.

Supponiamo ora di disporre di osservazioni sui numeri dei *q* tipi di sinistri riportati dai *p* rischi in più periodi, allora la verosimiglianza è data da

$$\ell = \prod_{i,t,j} P\left(N_{it}^{(j)} = n_{it}^{(j)} \middle| q_{1j}, x_{it}^{(j)}\right) = \prod_{i,t,j} e^{-l_{it}^{(j)}} \frac{{l_{it}^{(j)}}^{n_{it}^{(j)}}}{n_{it}^{(j)}!} .$$

Passando alla log-verosimiglianza, si ha

$$l = \sum_{i,t,j} \left(-l_{it}^{(j)} + n_{it}^{(j)} \log l_{it}^{(j)} - \log\left(n_{it}^{(j)}!\right)\right).$$

Calcolando le derivate di *l* rispetto a q_{1j} per $j = 1,\ldots, q$, con semplici calcoli si verifica che riesce

$$\frac{\partial l}{\partial q_{1j}} = \sum_{i,t} \left(n_{it}^{(j)} x_{it}^{(j)} - l_{it}^{(j)} x_{it}^{(j)}\right), \qquad j = 1,\ldots, q .$$

Dalla condizione $\frac{\partial l}{\partial q_{1j}} = 0$, con $j = 1,\ldots, q$ si ricavano le equazioni

$$\sum_{i,t} \left(n_{it}^{(j)} - l_{it}^{(j)}\right) x_{it}'^{(j)} = 0, \qquad j = 1,\ldots, q,$$

le cui soluzioni sono le stime di massima verosimiglianza dei vettori q_{1j}, $j = 1,\ldots, q$.

[1.4]

Passando ora al modello con eterogeneità, supponiamo che, per ogni *i*, vi sia un parametro aleatorio vettoriale $U_i = (U_{i1},...,U_{iq})$ non dipendente dal tempo e che per ogni $u_i = (u_{i1},...,u_{iq})$ determinazione possibile di U_i si abbia

- $N_{it}^{(j)} | U_i = u_i \quad j = 1,...,q$, $t = 1,...,T_i$, stocasticamente indipendenti,
- $N_{it}^{(j)} | U_i = u_i \sim P(l_{it}^{(j)} w_{ij})$, $l_{it}^{(j)} = \exp(x_{it}^{(j)\prime} q_{1j})$, $w_{ij} = \exp(u_{ij})$,
- $U_i \sim N(0,V)$ con V matrice di varianze e covarianze $V_{jk} = \mathrm{cov}(U_{ij}, U_{ik})$.

Inoltre, i processi relativi ai diversi individui siano stocasticamente indipendenti. I parametri del modello con eterogeneità sono dunque $(q_{1j})_{j=1,...,q}$ e V.

Osserviamo che l'ipotesi di indipendenza dei numeri aleatori $N_{it}^{(j)} | U_i = u_i$ al variare di j richiede che i numeri aleatori $N_{it}^{(j)}$ siano opportunamente definiti.

A titolo di esempio consideriamo nelle assicurazioni RCA, il caso della distinzione tra danni a cose e danni alle persone. Se indichiamo, per un individuo e per un periodo fissati, con N_A il numero di sinistri con danni a cose e con N_B il numero di sinistri con danni a persone, è evidente che non è ragionevole supporre indipendenza tra questi due numeri aleatori in quanto in un sinistro si possono provocare danni sia a cose che a persone. Possiamo però considerare i seguenti numeri aleatori:

- $N_{A \cap B}$: numero di sinistri con danni a cose e a persone
- N_{A-B} : numero di sinistri con danni solo a cose
- N_{B-A} : numero di sinistri con danni solo a persone

Per questi tre numeri aleatori può essere formulata l'ipotesi di indipendenza stocastica. Osserviamo che riesce

$$N_A = N_{A \cap B} + N_{A-B} \text{ e } N_B = N_{A \cap B} + N_{B-A}.$$

1.5 MODELLI PER IL COSTO DEI SINISTRI

In questo paragrafo vengono presentati due modelli adatti a descrivere il costo dei sinistri introducendo una dipendenza dalle caratteristiche tariffarie tramite le componenti di regressione. Si considerano le famiglie di distribuzioni gamma e log-normale. Tali distribuzioni sono indicizzate con due parametri: uno di scala, funzione delle componenti di regressione e uno di forma. Della componente di eterogeneità si tiene conto nel parametro di scala.

Fissato l'individuo i e il periodo t, indichiamo con N_{it} il numero di sinistri riportati e con C_{itj} il costo del j-esimo sinistro.

Descriviamo prima il modello base.

Con riferimento ai processi $\{N_{it}, C_{it1}, C_{it2}, ...\}$, si suppone che per ogni $n_{i1}, ..., n_{iT_i}$

- condizionatamente a $N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i}$ i vettori aleatori $(C_{i11}, ..., C_{i1n_{i1}}), ..., (C_{iT_i1}, ..., C_{iT_in_{iT_i}})$ siano stocasticamente indipendenti,
- per ogni t, $(C_{it1}, ..., C_{itn_{it}}) \big| N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i} =_d (C_{it1}, ..., C_{itn_{it}}) \big| N_{it} = n_{it}$,
- per ogni t, $C_{it1} \big| N_{it} = n_{it}, ..., C_{itn_{it}} \big| N_{it} = n_{it}$ sono i.i.d.

Inoltre si assume l'indipendenza tra i processi relativi ai diversi individui del portafoglio.

Consideriamo ora i due casi citati per la distribuzione di $C_{itj} \big| N_{it} = n_{it}$.

Si assuma per i costi una distribuzione gamma,

$$C_{itj} \big| N_{it} = n_{it} \sim gamma(d, b_{it}),\ b_{it} = \exp(x'_{it} b),\ i = 1, ..., p;\ t = 1, 2, ...;\ j = 1, ..., n_{it};$$

o, equivalentemente, $b_{it} C_{itj} \big| N_{it} = n_{it} \sim gamma(d, 1)$. La funzione di densità dei costi nel modello base è dunque

$$f(y) = \frac{b_{it}^{d}}{\Gamma(d)} y^{d-1} e^{-b_{it} y}\ ;\ b_{it} = \exp(x_{it}' b)$$

dove il coefficiente b_{it}, parametro di scala, è funzione del vettore x_{it} delle caratteristiche tariffarie dell'individuo i nel periodo t. I parametri del modello base sono dunque $q_1 = (b, d)$.

Nell'ipotesi in cui si disponga delle osservazioni sui costi dei sinistri riportati dagli individui si può ricavare la stima di massima verosimiglianza dei parametri.

Fissato l'individuo i, la verosimiglianza è data da:

$$\ell_i = \prod_{t,j} \frac{b_{it}^{d}}{\Gamma(d)} c_{itj}^{\ d-1} e^{-b_{it} c_{itj}}\ .$$

Poiché tra le ipotesi vi è quella dell'indipendenza per le quantità aleatorie che fanno riferimento ai singoli individui, si ottiene la verosimiglianza generale:

$$\ell = \prod_{i} \ell_i = \prod_{i,t,j} \frac{b_{it}^{d}}{\Gamma(d)} c_{itj}^{\ d-1} e^{-b_{it} c_{itj}}\ .$$

Passando alla log-verosimiglianza si ha:

$$l = \log \ell = \sum_{i,t,j} \left[d \log b_{it} + (d-1) \log c_{itj} - b_{it} c_{itj} - \log(\Gamma(d)) \right]$$

Derivando rispetto a b e a d si ottiene:

$$\frac{\partial l}{\partial b} = \frac{\partial l}{\partial b_{it}} \frac{\partial b_{it}}{\partial b} = \sum_{i,t,j} \left[\frac{d}{b_{it}} - c_{itj} \right] e^{x_{it} b} x_{it} = \sum_{i,t,j} \left[\frac{d}{b_{it}} - c_{itj} \right] b_{it} x_{it}\ , \qquad (1.5.1)$$

$$\frac{\partial l}{\partial d} = \sum_{i,t,j} \left[\log b_{it} + \log c_{itj} - \frac{\Gamma'(d)}{\Gamma(d)} \right].$$

Posto $\frac{\partial l}{\partial b} = 0$, $\frac{\partial l}{\partial d} = 0$ si ricavano le stime di massima verosimiglianza $\hat{b}$ e $\hat{d}$. Una volta ottenute tali stime, si può ricavare la stima del costo atteso per sinistro per l'assicurato i nel periodo t. Nel modello a priori dunque si ha:

[1.5]

$$\hat{c}_{it} = \frac{\hat{d}}{\hat{b}_{it}} = \frac{\hat{d}}{\exp(x'_{it}\hat{b})}.$$

Sostituendo i $\hat{c}_{it}$ nella condizione $\frac{\partial l}{\partial b} = 0$, si ha:

$$\sum_{i,t,j}\left[\frac{\hat{d}}{\hat{b}_{it}} - c_{itj}\right]\hat{b}_{it}x_{it} = \sum_{i,t}\sum_{j=1}^{n_{it}}\left[\hat{c}_{it} - c_{itj}\right]\frac{\hat{d}}{\hat{c}_{it}}x_{it} = \sum_{i,t}\sum_{j=1}^{n_{it}}\left[\frac{\hat{c}_{it} - c_{itj}}{\hat{c}_{it}}\right]\hat{d}x_{it} = 0.$$

Tale condizione è equivalente alla

$$\sum_{i,t}\left(\sum_{j=1}^{n_{it}}\left(1 - \frac{c_{itj}}{\hat{c}_{it}}\right)\right)x_{it} = \sum_{i,t} cres_{it}\,x_{it} = 0,$$

dove abbiamo posto $cres_{it} = \sum_{j=1}^{n_{it}}\left(1 - \frac{c_{itj}}{\hat{c}_{it}}\right)$ che rappresenta la somma degli scarti relativi tra valore stimato e valore osservato, rispetto al valore stimato, dei risarcimenti dei sinistri riportati dall'i-esimo assicurato nel t-esimo periodo. L'equazione di verosimiglianza per β può così essere interpretata come una relazione di ortogonalità tra le componenti di regressione e i residui $cres_{it}$.

In luogo della distribuzione gamma, per i costi, si può assumere una distribuzione di tipo log-normale. Ovvero si suppone che sia

$\log C_{itj} \,|\, N_{it} = n_{it} \sim N(m_{it}, s^2)$, con $m_{it} = x'_{it}b$.

In questo caso, allora, la funzione di densità del risarcimento per sinistro del modello base è:

$$f(y) = \frac{1}{\sqrt{2p}\,s\,y}\exp\left[-\frac{1}{2s^2}(\log y - m_{it})^2\right].$$[4]

[4] Ricordiamo che dato un numero aleatorio X, se $\log X$ ha distribuzione normale di media m e varianza s^2, X ha distribuzione log-normale con momento n-esimo

I parametri del modello sono dunque quelli del vettore $q_1 = (b, s^2)$.

Disponendo delle osservazioni sui costi dei sinistri riportati dagli individui si può ricavare la verosimiglianza:

$$\ell = \prod_{i,t,j} \frac{1}{\sqrt{2p}\, s\, c_{itj}} \exp\left[-\frac{1}{2s^2} (\log c_{itj} - m_{it})^2 \right]$$

e quindi la log-verosimiglianza:

$$l = \sum_{i,t,j} \left(-\log c_{itj} - \log s - \frac{1}{2s^2} (\log c_{itj} - m_{it})^2 - \log \sqrt{2p} \right).$$

Calcolando le derivate parziale rispetto a b e s^2 si ottiene:

$$\frac{\partial l}{\partial b} = \sum_{i,t} \frac{\partial l}{\partial m_{it}} \frac{\partial m_{it}}{\partial b} = \sum_{i,t,j} \frac{1}{s^2} (\log c_{itj} - m_{it}) x_{it} = \frac{1}{s^2} \sum_{i,t,j} (\log c_{itj} - m_{it}) x_{it} =$$
$$\sum_{i,t} \sum_{j} (\log c_{itj} - x_{it} b) x_{it} = \sum_{i,t} lcres_{it}\, x_{it},$$

$$\frac{\partial l}{\partial s^2} = \sum_{i,t,j} \left(-\frac{1}{2s^2} + \frac{1}{2s^4} (\log c_{itj} - m_{it})^2 \right),$$

dove si è posto $lcres_{it} = \sum_{j} (\log c_{itj} - m_{it})$.

Uguagliando a zero le derivate parziali si ricavano le stime di massima verosimiglianza cercate.

Si osservi che anche in questo caso, la condizione $\frac{\partial l}{\partial b} = 0$ ovvero la $\sum_{i,t} lcres_{it}\, x_{it} = 0$ si può leggere come una relazione di ortogonalità tra le componenti di regressione e i residui convenientemente definiti.

$E\left[X^n \right] = E\left[e^{n \log X} \right] = m_X(n) = \exp(nm + \frac{n^2}{2} s^2)$, dove $m_X(\cdot)$ indica la funzione generica dei momenti della distribuzione di X.

Introduciamo ora il modello con eterogeneità. Indicato con U_{ci} il parametro aleatorio associato all'individuo i, (il pedice c è introdotto per sottolineare che U_{ci} ha in particolare effetto sulla distribuzione dei costi) si suppone che per ogni $n_{i1},...,n_{iT_i},u$,

- condizionatamente a $(N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i},U_{ci}=u)$ i vettori aleatori $(C_{i11},...,C_{i1n_{i1}}),...,(C_{iT_i1},...,C_{iT_in_{iT_i}})$ siano stocasticamente indipendenti,
- per ogni t, $(C_{it1},...,C_{itn_{it}})\big|(N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i},U_{ci}=u)=_d$

$(C_{it1},...,C_{itn_{it}})\big|(N_{it}=n_{it},U_{ci}=u)$,

- per ogni t, $C_{it1}\big|(N_{it}=n_{it},U_{ci}=u),...,C_{itn_{it}}\big|(N_{it}=n_{it},U_{ci}=u)$ sono i.i.d..

Si assume ancora l'indipendenza tra i processi relativi ai diversi individui presenti nel portafoglio.

Si osservi che le attuali ipotesi ricalcano quelle del modello base che ora valgono tenendo conto anche del condizionamento al valore del parametro di rischio.

Ora precisando il tipo di distribuzione per i costi, nel caso gamma si richiede che

- $C_{itj}\,|N_{it}=n_{it},U_{ci}=u\sim gamma(d,b_{it}u)$, $b_{it}=x_{it}'b$,
- $U_{ci}\,\big|(N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i})=_d U_{ci}\,\big|(N_{it}=n_{it})=_d U_{ci}\sim gamma(d,d)$.

Nel caso log-normale

- $\log C_{itj}\,\big|(N_{it}=n_{it},U_{ci}=u)\sim N(m_{it}+u,s^2)$ $m_{it}=x_{it}'b$,
- $U_{ci}\,\big|(N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i})=_d U_{ci}\,\big|(N_{it}=n_{it})=_d U_{ci}\sim N(0,s_U^2)$.

Osserviamo che i parametri aleatori sono i.i.d. al variare di i.

1.6 MODELLI PER IL PREMIO PURO

Si indichi con $Y_i = (N_{it}, C_{it1}, C_{it2}, ...)_{t=1,...,T_i}$, il processo di osservazione dell'*i*-esimo assicurato. Fissato t, il risarcimento globale è $Z_{it} = \sum_{j=1}^{N_{it}} C_{itj}$ dove N_{it} è il numero aleatorio di sinistri nel periodo t e C_{itj} è il costo per il j-esimo sinistro.

In linea con le ipotesi introdotte per il modello base sui processi dei numeri dei sinistri e quelli dei costi, nel modello base si assumono le seguenti ipotesi:

- $N_{i1}, ..., N_{iT_i}$ stocasticamente indipendenti,

  per ogni $n_{i1}, ..., n_{iT_i}$,

- condizionatamente a $N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i}$ i vettori aleatori $(C_{i11}, ..., C_{i1n_{i1}}), ..., (C_{iT_i1}, ..., C_{iT_in_{iT_i}})$ siano stocasticamente indipendenti,
- per ogni t, $(C_{it1}, ..., C_{itn_{it}}) \big| (N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i}) =_d (C_{it1}, ..., C_{itn_{it}}) \big| N_{it} = n_{it}$,
- per ogni t, $C_{it1} \big| N_{it} = n_{it}, ..., C_{itn_{it}} \big| N_{it} = n_{it}$ sono i.i.d.,
- $N_{it} \sim P(l_{it})$, $l_{it} = \exp(x_{it}' b)$,
- $C_{itj} \big| N_{it} = n_{it} \sim gamma(d, b_{it})$, con $b_{it} = \exp(x_{it}' a)$,

  oppure

  $\log C_{itj} \big| N_{it} = n_{it} \sim N(m_{it}, s^2)$, con $m_{it} = x_{it}' a$.

Inoltre si suppone l'indipendenza tra i processi relativi ai diversi individui presenti nel portafoglio.

I parametri da stimare sono dunque i vettori a, b e lo scalare d oppure s^2.

Supponiamo ora di disporre delle osservazioni relative ai diversi individui, possiamo considerare la verosimiglianza che descriviamo solo nel caso dei costi log-normali.

Posto $E_i = \left(N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i} \right)$ e $H_i = \left(C_{i11} = c_{i11}, ..., C_{iT_i n_{iT_i}} = c_{iT_i n_{iT_i}} \right)$, si ha

$$\ell = \prod_i \left[P(E_i) \, f_{H_i|E_i}(c_{i11}, ..., c_{iT_i n_{iT_i}}) \right] = \prod_{i,t,j} \left[P(N_{it} = n_{it}) \, f_{C_{itj}|N_{it}=n_{it}}(c_{itj}) \right] =$$

$$= \prod_{i,t,j} \left[e^{-l_{it}} \frac{(l_{it})^{n_{it}}}{n_{it}!} \frac{1}{\sqrt{2p}\, s\, c_{itj}} e^{-\frac{1}{2s^2}\left(\log c_{itj} - m_{it}\right)^2} \right].$$

Quindi per la log-verosimiglianza si ha

$$l = \sum_{i,t,j} \left(n_{it} \log l_{it} - l_{it} - \log(n_{it}!) - \frac{\left(\log c_{itj} - m_{it}\right)^2}{2s^2} - \log s - \log c_{itj} - \log \sqrt{2p} \right),$$

dove, lo ricordiamo, $l_{it} = \exp(x_{it}' b)$, $m_{it} = x_{it}' a$.

Quindi la log-verosimiglianza è funzione di b, a e s^2.

Sottolineiamo anche che abbiamo considerato le stesse componenti di regressione per i numeri ed i costi di sinistri. Si potrebbe anche pensare a due vettori diversi di componenti di regressione, uno che abbia rilevanza per la descrizione probabilistica del numero di sinistri, l'altro per i costi (si veda Pinquet (2001b).

Le derivate della log-verosimiglianza rispetto ai parametri sono

$$\frac{\partial l}{\partial b} = \sum_{i,t} \frac{\partial l}{\partial l_{it}} \frac{\partial l_{it}}{\partial b} = \sum_{i,t} \left[n_{it} \frac{1}{l_{it}} - 1 \right] l_{it} x_{it} = \sum_{i,t} (n_{it} - l_{it}) x_{it},$$

$$\frac{\partial l}{\partial a} = \sum_{i,t} \frac{\partial l}{\partial m_{it}} \frac{\partial m_{it}}{\partial a} = \sum_{i,t,j} \frac{1}{s^2} \left(\log c_{itj} - m_{it} \right) x_{it},$$

$$\frac{\partial l}{\partial s^2} = \sum_{i,t,j} \left[\frac{1}{2s^4} \left(\log c_{itj} - m_{it} \right)^2 - \frac{1}{2s^2} \right]. \qquad (1.6.1)$$

Nel modello con eterogeneità occorre pensare a due componenti. Con riferimento al processo per il numero dei sinistri, fissato l'individuo i si considera il numero aleatorio U_{ni} che descrive l'eterogeneità connessa al numero di sinistri.

Per quel che riguarda il processo dei costi si considera invece il numero aleatorio U_{ci} che rappresenta l'eterogeneità collegata ai risarcimenti.

Se i random effects sono supposti indipendenti, il coefficiente bonus-malus per il premio puro è pari al prodotto dei coefficienti collegati alla frequenza e al costo atteso per sinistro. Tuttavia, si può pensare che il comportamento dell'assicurato influenzi in modo analogo i due random effects e quindi accogliere un'ipotesi di correlazione positiva tra essi. In questo caso occorre specificare una distribuzione congiunta per i random effects relativi al numero e al costo dei sinistri.

Descriviamo in modo dettagliato le ipotesi che si assumono nel modello per tenere conto anche di questo aspetto.

Si suppone che per ogni u_{ni}, u_{ci}

- condizionatamente a $U_{ni}=u_{ni}$, $U_{ci}=u_{ci}$, i numeri aleatori $N_{i1},...,N_{iT_i}$ siano stocasticamente indipendenti,

per ogni $n_{i1},...,n_{iT_i}$,

- condizionatamente a $N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i}$, $U_{ni}=u_{ni}$, $U_{ci}=u_{ci}$, i vettori aleatori $(C_{i11},...,C_{i1n_{i1}}),...,(C_{iT_i1},...,C_{iT_in_{iT_i}})$ siano stocasticamente indipendenti,
- per ogni t, $(C_{it1},...,C_{itn_{it}})\big|(N_{i1}=n_{i1},...,N_{iT_i}=n_{iT_i},U_{ni}=u_{ni},U_{ci}=u_{ci})=_d (C_{it1},...,C_{itn_{it}})\big|(N_{it}=n_{it},U_{ni}=u_{ni},U_{ci}=u_{ci})$,
- per ogni t, $C_{it1}\big|(N_{it}=n_{it},U_{ni}=u_{ni},U_{ci}=u_{ci}),...,C_{itn_{it}}\big|(N_{it}=n_{it},U_{ni}=u_{ni},U_{ci}=u_{ci})$ sono i.i.d.,
- $N_{it}\big|(U_{ni}=u_{ni},U_{ci}=u_{ci})=_d N_{it}\big|U_{ni}=u_{ni}\sim P(l_{it}e^{u_{ni}})$, $l_{it}=\exp(x_{it}'b)$,
- $\log C_{itj}\big|(N_{it}=n_{it},U_{ni}=u_{ni},U_{ci}=u_{ci})=_d \log C_{itj}\big|(N_{it}=n_{it},U_{ci}=u_{ci})$, $\log C_{itj}\big|(N_{it}=n_{it},U_{ci}=u_{ci})\sim N(m_{it}+u_{ci},s^2)$, con $m_{it}=x_{it}'a$.

Posto $U_i=\begin{pmatrix}U_{ni}\\U_{ci}\end{pmatrix}$

- $U_i \left| (N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i}) =_d U_i \right| (N_{it} = n_{it}) =_d U_i \sim N(0,V)$.

Si assume ancora l'indipendenza tra i processi relativi ai diversi individui presenti nel portafoglio.

Dunque si assume che nel modello con eterogeneità U_{ni} e U_{ci} seguano una distribuzione normale bivariata con speranza matematica nulla e matrice varianze-covarianze data da:

$$V = \begin{pmatrix} V_{nn} & V_{nc} \\ V_{nc} & V_{cc} \end{pmatrix}.$$

Ricordiamo che se U_i ha distribuzione normale bivariata, allora le due distribuzioni marginali sono normali. Si ha $U_{ni} \sim N(0,V_{nn})$ e $U_{ci} \sim N(0,V_{cc})$.

I parametri del modello sono:

$$q_1 = \begin{pmatrix} a \\ b \\ s^2 \end{pmatrix}; \; q_2 = \begin{pmatrix} V_{nn} \\ V_{cn} \\ V_{cc} \end{pmatrix}.$$

In questo caso Θ_2 è il cono delle matrici semi-definite positive, incluso nello spazio delle matrici simmetriche di dimensione 2 che è identificato da $\mathbb{R}^3$.

CAPITOLO 2

LA REVISIONE BAYESIANA E LA CREDIBILITÀ LINEARE

2.1 INTRODUZIONE.

Nel capitolo precedente si è detto che, nella tariffazione a priori, si opera una ripartizione degli assicurati in classi di rischio sulla base di caratteristiche tariffarie rilevabili a priori e si attribuisce lo stesso premio, detto *premio collettivo*, a tutti i rischi che appartengono ad una determinata classe.

Poiché all'interno di ogni classe tariffaria rimane notevole eterogeneità nei comportamenti degli assicurati e quindi nella sinistrosità, è diffuso il ricorso alla tariffazione in base all'esperienza nella quale il premio di ciascun assicurato viene aggiustato, a posteriori, sulla base dell'osservazione dei sinistri effettivamente riportati. Si perviene così ad una differenziazione del premio all'interno delle classi con l'obiettivo di far pagare, alla lunga, un premio corrispondente alla effettiva sinistrosità.

Tipicamente tale differenziazione viene raggiunta tramite la costruzione di opportuni coefficienti, detti *coefficienti bonus-malus*, che moltiplicati al premio a priori forniscono il premio a posteriori.

I modelli con eterogeneità possono venire utilizzati a tale scopo. In particolare, se si dispone della distribuzione del parametro aleatorio si può seguire un approccio puramente bayesiano. In tale caso, la revisione si ottiene attraverso il calcolo del valore atteso degli elementi aleatori che descrivono il rischio rispetto alla distribuzione a posteriori, e il coefficiente bonus-malus è il

rapporto tra i valori attesi con riferimento alla distribuzione a posteriori e a priori, rispettivamente.

La revisione del premio può essere ottenuta anche tramite modelli di credibilità. L'approccio della credibilità lineare può essere visto come una sorta di revisione bayesiana semplificata; infatti in tale metodo non occorre specificare la distribuzione del parametro aleatorio *U*, ma basta assegnare i suoi primi due momenti. Per i processi degli elementi aleatori che descrivono la rischiosità, in questo caso, si parla di *modelli mistura semiparametrici*.

Con questo metodo, la tariffazione sulla base dell'esperienza si raggiunge attraverso una media ponderata del premio collettivo e di un riassunto dell'esperienza individuale dell'assicurato. Il peso assegnato a tale riassunto viene detto *coefficiente di credibilità*, facendo chiaramente riferimento alla credibilità che si può attribuire alla storia dell'assicurato.

In questo capitolo presentiamo i due approcci, bayesiano e di credibilità, e li applichiamo ai modelli con eterogeneità presentati nel capitolo precedente.

2.2 LA REVISIONE BAYESIANA

Si consideri un fissato assicurato. Siano $(Y_1, Y_2, ...)$ il processo degli enti aleatori che descrivono la sua rischiosità e $(x_1, x_2, ...)$ la successione di componenti di regressione, al variare del tempo. Fissato un periodo t, indichiamo con $\underline{Y}_t$ e $\underline{x}_t$ i vettori $(Y_1, ..., Y_t)$ e $(x_1, ..., x_t)$, rispettivamente.

Si osservi che $\underline{x}_t$ e $\underline{Y}_t$ sostituiscono i simboli x_i e Y_i utilizzati nel capitolo precedente. È stato soppresso l'indice *i* che designa il particolare individuo dal momento che nelle ipotesi che abbiamo introdotto e ai fini delle valutazioni che si vogliono ottenere gli assicurati possono essere considerati separatamente. È stato invece indicato esplicitamente il numero *t* di componenti dei due vettori.

Supponiamo ora che l'assicurato sia stato osservato per T periodi. L'obiettivo è quello di formulare una previsione per un numero aleatorio che

riassuma la rischiosità dell'assicurato per il periodo T+1, attraverso un modello con eterogeneità. L'epoca in cui viene effettuata la valutazione sia l'anno T.

Il numero aleatorio che riassume la rischiosità sia una funzione di Y_{T+1} che indichiamo con $q(Y_{T+1})$. Indichiamo poi con R_{T+1} il valore atteso di $q(Y_{T+1})$, $R_{T+1} = E[q(Y_{T+1})]$. Per esempio, se Y_{T+1} è la sequenza del numero e del costo dei sinistri per il periodo T+1, $Y_{T+1} = (N_{T+1}, C_{T+11,} C_{T+12}, ...)$, e $q(Y_{T+1})$ è il risarcimento globale, R_{T+1} è il premio puro.

Gli elementi che caratterizzano il rischio, coinvolti nella valutazione sono quelli dei vettori $\underline{x}_{T+1}$ e Y_{T+1}. Accanto a $\underline{x}_{T+1}$ e $\underline{Y}_{T+1}$ consideriamo anche una componente di eterogeneità rappresentata, in generale, da un vettore $\underline{U}_{T+1} = (U_1, U_2, ..., U_{T+1})$, dove U_t è il valore del parametro di rischio relativo al periodo t. Se non vi è dipendenza dal tempo delle componenti di eterogeneità, il precedente vettore si riduce ad uno scalare.

Le ipotesi che regolano i legami probabilistici tra gli elementi aleatori del processo $Y_1, ..., Y_{T+1}, U_1, ..., U_{T+1}$ siano in linea con quelle descritte in dettaglio nei modelli considerati nel Capitolo 1. In particolare, riesca

- per ogni $t \le T+1$, $Y_t \,|\, \underline{U}_{T+1} = (u_1, ..., u_{T+1}) =_d Y_t \,|\, U_t = u_t$ e la comune distribuzione dipenda dal parametro q_1 e dalle componenti di regressione relative al periodo *t*, x_t;
- $Y_1 \,|\, \underline{U}_{T+1} = (u_1, ..., u_{T+1}), \ldots, Y_{T+1} \,|\, \underline{U}_{T+1} = (u_1, ..., u_{T+1})$ stocasticamente indipendenti.

Dunque la distribuzione di $Y_t | U_t = u_t$ dipende da q_1, x_t e u_t (per convenienza, indichiamo le densità marginali per ciascun periodo con lo stesso simbolo, ℓ^*, che nel Paragrafo 1.2 è stato usato per indicare la distribuzione congiunta per l'intera sequenza di periodi): $\ell^*(y_t | q_1, x_t, u_t)$.

Supponiamo che per il tipo di rischio cui siamo interessati riesca $E[q(Y_t) | q_1, x_t, u_t] = h_{q_1}(x_t) g(u_t)$, per ogni t, con h_{q_1} e g funzioni a valori reali. Ciò

accade, come mostreremo in dettaglio nel prossimo paragrafo, per i modelli relativi al numero dei sinistri, al costo di un sinistro e al risarcimento globale, presentati nel Capitolo 1.

Una previsione per il rischio, per il periodo T+1, è rappresentata dalla speranza matematica calcolata con la distribuzione a posteriori di Y_{T+1}:

$$E[q(Y_{T+1})\,|\,\underline{x}_{T+1}, \underline{Y}_T = (y_1,...,y_T)].$$

Per semplicità di scrittura indichiamo con H_T l'evento $\underline{Y}_T = (y_1,...,y_T)$. Per la proprietà di disintegrabilità della speranza matematica si ha

$$E[q(Y_{T+1})\,|\,\underline{x}_{T+1}, H_T] =$$
$$= \int E[q(Y_{T+1})\,|\,\underline{x}_{T+1}, H_T, \underline{U}_{T+1} = (u_1,...,u_{T+1})]dF_{\underline{U}_{T+1}|\underline{x}_{T+1},H_T}(u_1,...,u_{T+1}). \qquad (2.2.1)$$

In forza delle ipotesi sopra richiamate, il precedente integrale è pari a

$$\int E[q(Y_{T+1})|q_1, x_{T+1}, u_{T+1}]dF_{\underline{U}_{T+1}|\underline{x}_{T+1},H_T}(u_1,...,u_{T+1}) =$$
$$= h_{q_1}(x_{T+1})\int g(u_{T+1})dF_{\underline{U}_{T+1}|\underline{x}_{T+1},H_T}(u_1,...,u_{T+1}) = h_{q_1}(x_{T+1})E[g(U_{T+1})|\underline{x}_{T+1}, H_T]. \quad (2.2.2)$$

Possiamo ancora sviluppare la speranza matematica di condizionata di $g(U_{T+1})$ che compare all'ultimo membro. Infatti, se indichiamo con $\ell^*(y_t|q_1, x_t, u_t)$ la densità di $Y_t|U_t = u_t$, per la formula di Bayes, la densità di $\underline{U}_{T+1}|\underline{x}_{T+1}, H_T$ è

$$f_{\underline{U}_{T+1}|\underline{x}_{T+1},H_T}(u_1,...,u_{T+1}) = \frac{f_{\underline{U}_{T+1}}(u_1,...,u_{T+1})\prod_{t=1}^{T}\ell^*(y_t|q_1, x_t, u_t)}{E_{q_2}\left[\prod_{t=1}^{T}\ell^*(y_t|q_1, x_t, U_t)\right]}.$$

Allora, si ha

[2.2]

$$E[g(U_{T+1})|\underline{x}_{T+1},H_T]=\int g(u_{T+1})\frac{\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,u_t)}{E_{q_2}\left[\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right]}f_{\underline{U}_{T+1}}(u_1,...,u_{T+1})du_1du_2...du_{T+1}=$$

$$=\frac{E_{q_2}\left[g(U_{T+1})\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right]}{E_{q_2}\left[\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right]}.$$

La precedente relazione mette anche in evidenza che la speranza matematica a primo membro dipende da entrambi i parametri q_1 e q_2 e dalle componenti di regressione del vettore $\underline{x}_T$ mentre non dipende da x_{T+1}. Per questo motivo poniamo $E[g(U_{T+1})|\underline{x}_{T+1},H_T]=E_q[g(U_{T+1})|\underline{x}_T,H_T]$.

Dalle (2.2.1), (2.2.2) si ha dunque

$$E[q(Y_{T+1})|\ \underline{x}_{T+1},H_T]=h_{q_1}(x_{T+1})E_q[g(U_{T+1})|\underline{x}_T,H_T]. \qquad (2.2.3)$$

Osserviamo che è facile verificare che

$$E_q[g(U_{T+1})|\underline{x}_T,H_T]=\frac{E_{q_2}\left[g(U_{T+1})\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right]}{E_{q_2}\left[\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right]}$$

è la soluzione del seguente problema di minimo

$$\min_a E_{q_2}\left([g(U_{T+1})-a]^2\prod_{t=1}^{T}\ell^*(y_t|q_1,x_t,U_t)\right).$$

Se a secondo membro della (2.2.3) si sostituiscono q_1, q_2 con loro stime, si ottiene una stima del premio a posteriori per il periodo T+1 che indichiamo con $\hat{R}_{T+1}^T$. Si ha

$$\hat{R}_{T+1}^T=h_{\hat{q}_1}(x_{T+1})E_{\hat{q}}[g(U_{T+1})|\underline{x}_T,H_T].$$

Tale espressione può essere scritta come:

$$\left(h_{\hat{q}_1}(x_{T+1})E_{\hat{q}_2}[g(U_{T+1})]\right)\cdot\frac{E_{\hat{q}}[g(U_{T+1})|x_1,...,x_T;y_1,...,y_T]}{E_{\hat{q}_2}[g(U_{T+1})]}. \qquad (2.2.4)$$

Osserviamo che il primo fattore è il premio a priori per il periodo T+1, basato sui fattori tariffari del periodo corrente, si ha infatti

$$E[q(Y_{T+1})|x_{T+1}]=E_{q_2}\left[E[q(Y_{T+1})|q_1,x_{T+1},U_{T+1}]\right]=E_{q_2}\left[h_{q_1}(x_{T+1})g(U_{T+1})\right].$$

Mentre il secondo termine è un coefficiente bonus-malus che stima il rapporto di due valori attesi della stessa variabile, calcolati con le distribuzioni a posteriori e a priori, rispettivamente.

Si osservi, infine, che nei procedimenti bayesiani si richiede che la distribuzione mistura sia data a priori, mentre, in questo contesto, tale distribuzione viene stimata dai dati.

2.3 APPLICAZIONE AI MODELLI CON ETEROGENEITÀ

In questo paragrafo calcoliamo la speranza matematica a posteriori per i modelli presentati nel Capitolo 1. Come nel Paragrafo 2.2, per semplicità, consideriamo un fissato individuo e non indichiamo esplicitamente l'indice individuale.

Esempio 1

Iniziamo dal modello binomiale negativo per il numero di sinistri descritto nel Paragrafo 1.3. La componente di eterogeneità è uno scalare U e si ha

$$N_t|U=u\sim P(l_t w),\ l_t=\exp(x_t'q_1),\ \text{w=u},$$

con $W=U\sim gamma(d,d)$.

Poiché

$$E(N_t|U=u) = l_t u\,,$$

con le notazioni utilizzate nel paragrafo precedente si ha:

$$h_{q_1}(x_t) = \exp(x_t' q_1)\,; \quad g(u) = u\,;$$

$$Var(U) = s^{\,2} = q_2 = \frac{1}{d}\,.$$

Poiché $E_{q_2}(U) = 1$, per ogni q_2, per quanto illustrato nel Paragrafo 1.3, Esempio 1, se $H_T = (N_1 = n_1,...,N_T = n_T)$, il coefficiente bonus-malus ricavato dalla (2.2.4) è:

$$E_{\hat{q}}\left[U\,|\,\underline{x}_T, H_T\right] = \hat{u}^{T+1} = \frac{1+\hat{s}^{\,2}\sum_{t=1}^{T} n_t}{1+\hat{s}^{\,2}\sum_{t=1}^{T}\hat{l}_t} \qquad (2.3.1)$$

Si osservi che se la stima $\hat{u}^{T+1} - 1$ è negativa, cioè se la somma dei residui $\sum_{t=1}^{T}(n_t - \hat{l}_t)$ riesce minore di zero, allora si ha un "bonus per la frequenza", nel senso che il numero atteso di sinistri, per il periodo T+1, calcolato in base all'esperienza di sinistrosità dell'assicurato è minore di quello calcolato a priori.

Poiché $N_1|U=u,...,N_t|U=u$ sono stocasticamente indipendenti e $N_t|U=u$ ha distribuzione di Poisson di parametro $\hat{l}_t u$, per la successione sussiste il teorema di Cebicev (è infatti immediato constatare che le varianze dei numeri aleatori del processo sono equilimitate ovvero esiste *M* tale che $\hat{l}_t u \le M$, per ogni *t*). Allora la successione delle medie aritmetiche soddisfa la legge debole dei grandi numeri, cioè la successione

$$\frac{\sum_{t=1}^{T} N_t|U=u}{T} - E\left(\frac{\sum_{t=1}^{T} N_t|U=u}{T}\right) = \frac{\sum_{t=1}^{T} N_t|U=u}{T} - u\frac{\sum_{t=1}^{T}\hat{l}_t}{T}\,,\ T = 1,2,...$$

converge in probabilità a zero per $T \to +\infty$.

Dal punto di vista pratico, questo risultato suggerisce che per T elevato il valore osservato di $\frac{\sum_{t=1}^{T} N_t | U = u}{T}$ è con probabilità elevata prossimo a $u\frac{\sum_{t=1}^{T} \hat{l}_t}{T}$. Quindi se T è elevato, nella (2.3.1), $\sum_{t=1}^{T} n_t$ è approssimabile con $\sum_{t=1}^{T} \hat{l}_t$ e pertanto $E_{\hat{q}}\left[U | \underline{x}_T, H_T\right]$ è approssimabile con *u*. Dunque per le distribuzioni del numero dei sinistri, l'eterogeneità che non è spiegata dai fattori tariffari è rivelata completamente attraverso il tempo.

Esempio 2

Nel caso del modello per i sinistri di diversi tipi trattato nel Paragrafo 1.4, la componente di eterogeneità è un vettore invariante al variare del tempo $U = (U_1,...,U_q)$ e si pone

$$N_t^{(j)} | U = u \sim P(l_t^{(j)} w_j),\ l_t^{(j)} = \exp(x_t^{(j)\prime} q_{1j}),\ w_j = \exp(u_j),\ j = 1,...,q;\ t = 1,2,...$$

Supponiamo inoltre che *U* abbia distribuzione normale multivariata, $U \sim N(0,V)$.

Si ha dunque

$$E(N_t^{(j)} | U = u) = l_t^{(j)} \exp(u_j),$$

pertanto, con le notazioni utilizzate nel paragrafo precedente, per la j-esima componente si ha:

$$\left(h_{q_1}(x_t)\right)_j = \exp(x_t^{(j)\prime} q_{1j});\quad \left(g(u)\right)_j = \exp(u_j).$$

Indicato con H_T l'evento $\underline{Y}_T = (n_1,...,n_T)$ con $n_t = (n_t^{(1)},...,n_t^{(q)})$, la distribuzione di *U* condizionata a $\underline{x}_T$ e H_T è data da:

[2.3]

$$f_{U|\underline{x}_T,H_T}(u) \propto f(u)\prod_{j=1}^{q}\prod_{t=1}^{T}\exp(-l_t^{(j)}e^{u_j})\frac{(l_t^{(j)}e^{u_j})^{n_t^{(j)}}}{n_t^{(j)}!} =$$

$$= kf(u)\exp\left(\sum_{j=1}^{q}\sum_{t=1}^{T}(-l_t^{(j)}e^{u_j}) + \sum_{j=1}^{q}\sum_{t=1}^{T}u_j n_t^{(j)}\right)$$

dove $f(u)$ indica la densità (a priori) di U e k è una costante rispetto ad u.

In particolare, ricordando che $\int f_{U|\underline{x}_T,H_T}(u)du = 1$, si ha

$$k = \frac{1}{\int f(u)\exp\left(-\sum_{j=1}^{q}\sum_{t=1}^{T}l_t^{(j)}e^{u_j} + \sum_{J=1}^{q}\sum_{t=1}^{T}u_j n_t^{(j)}\right)du}.$$

Dunque per i sinistri di tipo *h* il coefficiente di aggiornamento è dato da

$$\frac{E_{\hat{q}}[g(U_h)|x_1,...,x_T;n_1,...,n_T]}{E_{\hat{q}_2}[g(U_h)]} = \frac{\int\exp\left(u_h - \sum_{j=1}^{q}\sum_{t=1}^{T}l_t^{(j)}e^{u_j} + \sum_{j=1}^{q}\sum_{t=1}^{T}u_j n_t^{(j)}\right)g(u)du}{\int\exp\left(-\sum_{j=1}^{q}\sum_{t=1}^{T}l_t^{(j)}e^{u_j} + \sum_{j=1}^{q}\sum_{t=1}^{T}u_j n_t^{(j)}\right)g(u)du} =$$

$$= \frac{E_{\hat{q}_2}\left(\exp\left(U_h - \sum_{j=1}^{q}\sum_{t=1}^{T}l_t^{(j)}e^{U_j} + \sum_{j=1}^{q}\sum_{t=1}^{T}U_j n_t^{(j)}\right)\right)}{E_{\hat{q}_2}(e^{U_h})\cdot E_{\hat{q}_2}\left(\exp\left(-\sum_{j=1}^{q}\sum_{t=1}^{T}l_t^{(j)}e^{U_j} + \sum_{j=1}^{q}\sum_{t=1}^{T_j}U_j n_t^{(j)}\right)\right)} =$$

$$= \frac{E_{\hat{q}_2}\left(\exp\left(U_h - \sum_{j=1}^{q}l^{(j)}e^{U_j} + \sum_{j=1}^{q}U_j n^{(j)}\right)\right)}{E_{\hat{q}_2}(e^{U_h})\cdot E_{\hat{q}_2}\left(\exp\left(-\sum_{j=1}^{q}l^{(j)}e^{U_j} + \sum_{j=1}^{q}U_j n^{(j)}\right)\right)},$$

dove abbiamo posto $l^{(j)} = \sum_{t=1}^{T}l_t^{(j)}$ e $n^{(j)} = \sum_{t=1}^{T}n_t^{(j)}$.

Esempio 3

Passiamo ora al modello gamma-gamma per i costi dei sinistri.

Vogliamo ricavare il coefficiente bonus-malus per il costo atteso per sinistro con il criterio descritto nel Paragrafo 2.2. Dal momento che si intende realizzare tale obiettivo utilizzando solo il modello con eterogeneità per le distribuzioni dei costi, supponiamo che i parametri aleatori relativi al numero e ai costi dei sinistri siano indipendenti, e indichiamo con U_c quello relativo ai costi.

Con riferimento al primo modello con eterogeneità presentato nel Paragrafo 1.5, posto $E_T = (N_1 = n_1, ..., N_T = n_T)$ si ha $C_{tj} \,|\, E_T, U_c = u \sim gamma(d, b_t u)$, $b_t = x_t' b$, quindi $E(C_{tj} \,|\, E_T, U_c = u) = \frac{d}{b_t u}$, $h_{q_1}(x_t) = \frac{d}{x_t' b}$ e $g(u) = \frac{1}{u}$.

Inoltre riesce $U_c \,|\, E_T =_d U_c \sim gamma(d, d)$, quindi $Var(U_c)) = s_U^2 = \frac{1}{d}$. Richiediamo che riesca $d > 1$. Questa è una condizione necessaria affinché $\frac{1}{U_c}$ sia dotato di speranza matematica. È facile verificare che $E\left(\frac{1}{U_c}\right) = \frac{d}{d-1}$.

Posto H_T per l'evento prodotto logico di $(C_{11}, ..., C_{1n_1}, ..., C_{T1}, ..., C_{Tn_T}) = (c_{11}, ..., c_{1n_1}, ..., c_{T1}, ..., c_{Tn_T})$ e di E_T, dal teorema di Bayes la distribuzione a posteriori di $U_c \,|\, E_T$ visti $\underline{x}_T, H_T$ è:

$$f_{U|\underline{x}_T, H_T} \propto u^{d-1} \exp(-d\,u) u^{d \sum_t n_t} \exp\left(-u \sum_{t,j} b_t c_{tj} \right).$$

Una volta osservata la storia dell'assicurato dunque, la distribuzione a posteriori di U è una $gamma(d + d\sum_t n_t, d + \sum_{t,j} b_t c_{tj})$ e riesce

$$E_q\left[\frac{1}{U_c} \,|\, \underline{x}_T, H_T \right] = \frac{d + \sum_{t,j} b_t c_{tj}}{d - 1 + d \sum_t n_t}$$

Allora il coefficiente bonus-malus è:

[2.3]

$$\frac{E_{\hat{q}}\left[\frac{1}{U_c}\middle|\underline{x}_T, H_T\right]}{E_{\hat{q}_2}\left[\frac{1}{U_c}\right]} = \frac{\hat{d}+\sum_{t,j}\hat{b}_t c_{tj}}{\frac{\hat{d}-1+\hat{d}\sum_t n_t}{\hat{d}}} = \frac{(\hat{d}-1)(\hat{d}+\sum_{t,j}\hat{b}_t c_{tj})}{\hat{d}(\hat{d}-1+\hat{d}\sum_t n_t)} = \frac{\hat{h}+\frac{\hat{d}-1}{\hat{d}}\frac{\sum_{t,j}\hat{b}_t c_{tj}}{\hat{d}}}{\hat{h}+tn_T},$$

dove abbiamo posto $\hat{h} = \frac{\hat{d}-1}{\hat{d}}$ e $tn_T = \sum_t n_t$.

Osservato che $E_{\hat{q}}\left(C_{tj}\middle|N_t = n_t\right) = E_{\hat{q}_2}\left[\frac{\hat{d}}{\hat{b}_t U}\right] = \frac{\hat{d}}{\hat{b}_t}\frac{\hat{d}}{\hat{d}-1}$, si ha

$$\frac{E_{\hat{q}}\left[\frac{1}{U_c}\middle|\underline{x}_T, H_T\right]}{E_{\hat{q}_2}\left[\frac{1}{U_c}\right]} = \frac{\hat{h}+\sum_t n_t \frac{c_t}{\hat{c}_t}}{\hat{h}+tn_T}$$

dove $c_t = \frac{\sum_j c_{tj}}{n_t}$ rappresenta il valore osservato del costo medio per sinistro all'epoca t e $\hat{c}_t = \frac{\hat{d}}{\hat{b}_t}\frac{\hat{d}}{\hat{d}-1}$ rappresenta il valore atteso, stimato, dello stesso costo.

Indicando con $cres_T = \sum_t n_t\left(1-\frac{c_t}{\hat{c}_t}\right)$ la somma dei residui dei costi per tutto il periodo di osservazione, si avrà un "bonus per i costi" se il residuo $cres_T$ è positivo. Il bonus è dunque pari a

$$1-\frac{\hat{h}+\sum_t n_t \frac{c_t}{\hat{c}_t}}{\hat{h}+tn_T} = \frac{cres_T}{\hat{h}+tn_T}$$

Esempio 4

Presentiamo il coefficiente bonus-malus per il modello log-normale per il costo dei sinistri.

[2.3]

Utilizzando le notazioni introdotte nel Paragrafo 1.5 e posto $E_T = (N_1 = n_1, ..., N_T = n_T)$ si ha:

$$\log C_{tj} \,|\, E_T, U_c = u \sim N(m_t + u, s^2), \; m_t = x_t' b$$

$$U_c \,|\, E_T =_d U_c \sim N(0, s_U^2)$$

e quindi:

$$E(C_{tj} \,|\, E_T, U_c = u) = \exp(m_t + u + \frac{s^2}{2});$$

$$h_{q_1}(x_t) = \exp(m_t + \frac{s^2}{2}); \qquad g(u) = \exp(u).$$

Posto $tn_T = \sum_{t=1}^{T} n_t$, e indicato con H_T, come nell'Esempio 3, l'evento prodotto logico di E_T e $(C_{11}, ..., C_{1n_1}, ..., C_{T1}, ..., C_{Tn_T}) = (c_{11}, ..., c_{1n_1}, ..., c_{T1}, ..., c_{Tn_T})$, dal teorema di Bayes si ha:

$$f_{U_c | \underline{x}_T, H_T}(u) \propto f(u) \prod_{t,j} \frac{1}{\sqrt{2p}\, s\, c_{tj}} \exp\left(-\frac{1}{2s^2} \left(\log c_{tj} - m_t - u \right)^2 \right) \propto$$

$$\propto \exp\left[-\frac{u^2}{2} \left(\frac{1}{s_U^2} + \frac{tn_T}{s^2} \right) \right] \cdot \exp\left[\frac{u}{s^2} \sum_{t,j} (\log c_{tj} - m_t) \right] \propto$$

$$\propto \exp\left[-\frac{1}{2} \left(\frac{1}{s_U^2} + \frac{tn_T}{s^2} \right) \left(u^2 - 2 \frac{u}{s^2} \frac{\sum_{t,j} (\log c_{tj} - m_t)}{\frac{1}{s_U^2} + \frac{tn_T}{s^2}} \right) \right].$$

La distribuzione a posteriori è dunque una normale:

$$U_c \,|\, \underline{x}_T, H_T \sim N\left(\frac{\sum_{t,j} (\log c_{tj} - m_t)}{\frac{s^2}{s_U^2} + tn_T}, \frac{1}{\frac{1}{s_U^2} + \frac{tn_T}{s^2}} \right).$$

Conseguentemente, si ha

[2.3]

$$E\left[e^{U_c}\,\middle|\,\underline{x}_T, H_T\right] = \exp\left(\frac{\sum_{t,j}(\log c_{tj} - m_t)}{\frac{s^2}{s_U^2} + tn_T} + \frac{1}{2}\frac{1}{\frac{1}{s_U^2} + \frac{tn_T}{s^2}}\right).$$

Posto $lcres_T = \sum_{t,j}(\log c_{tj} - m_t)$, possiamo scrivere

$$E_{\hat{q}}\left[\exp(U_c)\,\middle|\,\underline{x}_T, H_T\right] = \exp\left(\frac{lcres_T + \frac{s^2}{2}}{\frac{s^2}{s_U^2} + tn_T}\right).$$

Ricordando che $E(e^{U_c}) = e^{\frac{s_U^2}{2}}$, con semplici passaggi si ottiene l'espressione del coefficiente bonus-malus:

$$\frac{E_{\hat{q}}\left[\exp(U_c)\,\middle|\,\underline{x}_T, H_T\right]}{E_{\hat{q}_2}\left[\exp(U_c)\right]} = \exp\left[\frac{lcres_T - \frac{tn_T \hat{s}_U^2}{2}}{\frac{\hat{s}^2}{\hat{s}_U^2} + tn_T}\right].$$

Osserviamo che poiché $E(\log C_{tj}\,|\,E_T) = m_t$, $lcres_T$ rappresenta la somma dei residui dei logaritmi dei costi.

Esempio 5

Concludiamo questo paragrafo con la valutazione dei coefficienti bonus-malus per il modello del premio puro descritto nel Paragrafo 1.6.

Con le notazioni adottate nel Capitolo 1, posto $Z_t = \sum_{j=1}^{N_t} C_{tj}$, e $E_T = (N_1 = n_1, ..., N_T = n_T)$ si ha:

$$N_t\,|U_n = u_n \sim P(l_t e^{u_n}),\ l_t = \exp(x_t' b)$$

$$\log C_{tj}\,|E_T, U_c = u_c \sim N(m_t + u_c, s^2),\ m_t = x_t' a$$

$$U = \begin{pmatrix} U_n \\ U_c \end{pmatrix} \sim N(0, V).$$

Dunque riesce

$$E\left[Z_t \middle| U_n = u_n, U_c = u_c \right] = l_t e^{m_t + \frac{1}{2}s^2} e^{u_n} e^{u_c};$$

$$h_{q_1}(x_t) = l_t \exp\left(m_t + \frac{1}{2}s^2 \right) \quad g(u) = \exp(u_n + u_c).$$

La speranza matematica $E_{\hat{q}_2}\left[e^{U_n + U_c} \right]$, può venire formulata in funzione dei parametri V_{nn}, V_{cn} e V_{cc} elementi della matrice *V*. Presentiamo qui di seguito l'espressione di questa speranza matematica, riportata in Pinquet (1997).

Si parta dalla *relazione di scomposizione di Cholesky* per la matrice *V*,

$$V = T_j T_j' = \begin{bmatrix} j_{nn}^2 & j_{nn} j_{cn} \\ j_{nn} j_{cn} & j_{cn}^2 + j_{cc}^2 \end{bmatrix},$$

dove $T_j = \begin{bmatrix} j_{nn} & 0 \\ j_{cn} & j_{cc} \end{bmatrix}$ è una matrice triangolare inferiore. Si ha dunque $j_{nn}^2 = V_{nn}$, $j_{nn} j_{cn} = V_{nc}$, $j_{cn}^2 + j_{cc}^2 = V_{cc}$.

È facile verificare che esiste un vettore con distribuzione normale bidimensionale standard, $S = \begin{pmatrix} S_n \\ S_c \end{pmatrix} \sim N(0, I_2)$, dove I_2 rappresenta la matrice identica di ordine due, tale che possiamo esprimere le componenti del vettore *U* tramite la relazione $T_j S = U$. Infatti, se vale questa condizione, allora riesce

$$U_n = j_{nn} S_n$$
$$U_c = j_{cn} S_n + j_{cc} S_c.$$

Si può facilmente verificare che il vettore S di componenti $S_n = \frac{U_n}{j_{nn}}$, $S_c = \frac{1}{j_{nn}}\left[U_c - \frac{j_{cn}}{j_{nn}}U_n\right]$ soddisfa le precedenti condizioni.

Da questa osservazione, poiché S_n, S_c sono stocasticamente indipendenti, si ricava

$$E\left[e^{U_n+U_c}\right] = E\left[e^{(j_{nn}+j_{cn})S_n + j_{cc}S_c}\right] = E\left[e^{(j_{nn}+j_{cn})S_n}\right]E\left[e^{j_{cc}S_c}\right].$$

Osservando ora che, nella precedente uguaglianza, i due fattori sono i valori della funzione generatrice dei momenti della normale standard in $(j_{nn} + j_{cn})$ e j_{cc}, rispettivamente, si ottiene (si veda nota a piè di pagina 4)

$$E\left[e^{U_n+U_c}\right] = e^{\frac{1}{2}(j_{nn}+j_{cn})^2} e^{\frac{1}{2}j_{cc}^2} = e^{\frac{1}{2}\left[(j_{nn}+j_{cn})^2 + j_{cc}^2\right]} = e^{\frac{1}{2}\left(j_{nn}^2 + 2j_{nn}j_{cn} + j_{cn}^2 + j_{cc}^2\right)} = e^{\frac{1}{2}(V_{nn}+2V_{cn}+V_{cc})}.$$

Ritornando ora al coefficiente bonus-malus per il premio puro, l'espressione della verosimiglianza dell'osservazione $\left(n_1,..,n_T,c_{11},...,c_{Tn_T}\right)$ è data da

$$\ell = \prod_t \left[\frac{\left(l_t e^{u_n}\right)^{n_t}}{n_t!} e^{-l_t e^{u_n}} \prod_j \frac{1}{\sqrt{2p}\, s\, c_{tj}} \exp\left(-\frac{1}{2s^2}\left(\log c_{tj} - m_t - u_c\right)^2 \right) \right] =$$

$$= k \exp\left(u_n \sum_t n_t \right) \exp\left(-e^{u_n} \sum_t l_t \right) \exp\left(\sum_t \sum_j -\frac{1}{2s^2}\left(-2u_c\left(\log c_{tj} - m_t\right) + u_c^2\right) \right) =$$

$$= k \exp\left(u_n tn_T\right) \exp\left(-e^{u_n} \sum_t l_t \right) \exp\left(-\frac{1}{2s^2}\left(-2u_c lcres_T + tn_T u_c^2\right) \right),$$

dove si è posto $tn_T = \sum_{t=1}^{T} n_t$, e $lcres_T = \sum_{t,j} (\log c_{tj} - m_t)$.

Posto

$$v_T = u_n tn_T - e^{u_n} \sum_t l_t - \frac{1}{2s^2}\left(-2u_c lcres_T + tn_T u_c^2\right),$$

riesce $\ell = k\,\mathrm{e}^{v_T}$, dove $k = \dfrac{1}{E\left(\mathrm{e}^{v_T}\right)}$ è una costante di normalizzazione.

Indicato con H_T, l'evento prodotto logico di $(C_{11},...,C_{1n_1},...,C_{T1},...,C_{Tn_T}) = (c_{11},...,c_{1n_1},...,c_{T1},...,c_{Tn_T})$ e di E_T, dal teorema di Bayes si ha:

$$f_{U|\underline{x}_T,H_T}(u) \propto f(u)e^{v_T}.$$

Allora

$$E_q\left[e^{U_n+U_c}\,|\underline{x}_T, H_T\right] = k\int e^{U_n+U_c+v_t} f(u)du.$$

Se indichiamo con $V_T = U_n tn_T - e^{U_n}\sum_t l_t - \dfrac{1}{2s^2}\left(-2U_c lcres_T + tn_T U_c^2\right)$, si ha che

$$k = \frac{1}{E_q(V_T)} \text{ e } E_q\left[e^{U_n+U_c}\,|\underline{x}_T, H_T\right] = \frac{E_{q_2}\left[e^{U_n+U_c+V_T}\right]}{E_{q_2}\left[e^{V_T}\right]}.$$

Allora l'espressione del coefficiente bonus malus è data da

$$\frac{E_{\hat{q}}\left[e^{U_n+U_c}\,|\underline{x}_T, H_T\right]}{E_{\hat{q}_2}\left[e^{U_n+U_c}\right]} = \frac{E_{\hat{q}_2}\left[e^{U_n+U_c+V_T}\right]}{E_{\hat{q}_2}\left[e^{V_T}\right]E_{\hat{q}_2}\left[e^{U_n+U_c}\right]}. \qquad (2.3.2)$$

Il coefficiente bonus-malus dunque dipende da $\sum_t \hat{l}_t$, dal numero di sinistri riportati dall'assicurato nei T periodi, tn_T, ed infine dalla somma dei residui del logaritmo dei costi dei sinistri. Inoltre viene calcolato attraverso la simulazione di S_n e S_c.

2.4 LA CREDIBILITÀ LINEARE

In questo paragrafo richiamiamo brevemente l'approccio della credibilità lineare (si veda ad es. Klugman, Panjer, Wilmot (1984)). In particolare presentiamo i modelli di Bühlmann e di Bühlmann-Straub.

Fissato un individuo della collettività, utilizzando la simbologia già adottata, indichiamo con $(Y_1, Y_2, ...)$ il processo di osservazione degli enti aleatori che descrivono la sua rischiosità. Introdotto un parametro aleatorio U, supponiamo che $Y_1 | U = u, ..., Y_T | U = u$ siano numeri aleatori stocasticamente indipendenti. Per quanto riguarda il parametro U,come abbiamo detto l'approccio della credibilità lineare non richiede di specificarne la distribuzione, ma soltanto di supporre l'esistenza finita dei primi due momenti.

L'obiettivo è quello di ottenere uno stimatore per Y_{T+1} basato su $\underline{Y}_T = (Y_1, ..., Y_T)$. L'approccio bayesiano lineare (o di credibilità) prevede di costruire tale stimatore attraverso la funzione lineare (affine) f soluzione del seguente problema di minimo

$$\min E\left[\left(Y_{T+1} - f(\underline{Y}_T) \right)^2 \right], \text{ con } f(\underline{Y}_T) = b_0 + \sum_t b_t Y_t .$$

Dunque lo stimatore di credibilità di Y_{T+1} è dato da

$$\tilde{Y}_{T+1} = b_0 + \sum_{t=1}^{T} b_t Y_t$$

dove $b_0, b_1, ..., b_T$ si ricavano determinando il minimo della funzione

$$Q(b_0, ..., b_T) = E\left[Y_{T+1} - \left(b_0 + \sum_{t=1}^{T} b_t Y_t \right) \right]^2 .$$

Tale minimo si ottiene risolvendo il sistema lineare

$$\begin{cases} \dfrac{\partial Q}{\partial b_t} = 0 , & t = 0, ..., T . \end{cases}$$

Come è ben noto, tale stima è equivalente al cosiddetto *sistema normale*:

$$\begin{cases} E(Y_{T+1}) - \left(b_0 + \sum_{t=1}^{T} b_t E(Y_t) \right) = 0 \\ \operatorname{cov}(Y_{T+1}, Y_t) = \sum_{h=1}^{T} b_h \operatorname{cov}(Y_h, Y_t), \qquad t = 1,...,T. \end{cases}$$

Posto $\tilde{Y}_{T+1} = b_0 + \sum_{t=1}^{T} b_t Y_t$, il sistema precedente può venire riscritto come

$$\begin{cases} E(Y_{T+1}) = E(\tilde{Y}_{T+1}) \\ \operatorname{cov}(Y_{T+1}, Y_t) = \operatorname{cov}(\tilde{Y}_{T+1}, Y_t), \qquad t = 1,...,T. \end{cases} \tag{2.4.1}$$

È facile verificare che se poniamo

$$Q_1(b_0,...,b_T) = E\left[E\left(Y_{T+1} \middle| U\right) - b_0 - \sum_{t=1}^{T} b_t Y_t \right]^2$$

e

$$Q_2(b_0,...,b_T) = E\left[E\left(Y_{T+1} \middle| Y_1,...,Y_T\right) - b_0 - \sum_{t=1}^{T} b_t Y_t \right]^2,$$

i due sistemi che si ottengono imponendo l'annullamento delle derivate parziali di Q_1 e Q_2 sono equivalenti al sistema normale. Pertanto, $\tilde{Y}_{T+1} = b_0 + \sum_{t=1}^{T} b_t Y_t$, con $b_0,...,b_T$ soluzioni del sistema normale, è il migliore stimatore lineare, nel senso della funzione di perdita quadratica non solo di Y_{T+1}, ma anche del *premio individuale* $E\left(Y_{T+1} \middle| U\right)$ e del *premio bayesiano* $E\left(Y_{T+1} \middle| Y_1,...,Y_T\right)$.

Modello di Bühlmann

In particolare se $Y_1 | U = u, Y_2 | U = u, ...$ sono anche identicamente distribuiti, il processo $(Y_1, Y_2, ...)$ è scambiabile. Lo stimatore che realizza il minimo deve necessariamente essere del tipo

$$\tilde{Y}_{T+1} = b_0 + b\bar{Y}_T \qquad \text{con } \bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \,.$$

Inoltre, come è noto, si ha

$$b_0 = E(Y_{T+1}) - bE(\bar{Y}_T) = m(1-b)\,;$$

$$b = \frac{\operatorname{cov}(Y_{T+1}, \bar{Y}_T)}{Var(\bar{Y}_T)}\,,$$

dove $m = E(Y_t)$ per ogni t.

Osserviamo che per la proprietà iterativa della speranza matematica, si ha

$$E(Y_t) = E\left[E(Y_t|U)\right] = E\left[m(U)\right] = m\,.$$

Se poniamo $v = E\left[Var(Y_t|U)\right]$, $w = Var\left[E(Y_t|U)\right] = Var(m(U))$, il coefficiente *b* può essere espresso tramite *v* e *w*. Infatti, si ha

$$Var(\bar{Y}_T) = E\left[Var(\bar{Y}_T|U)\right] + Var\left[E(\bar{Y}_T|U)\right] = E\left[\frac{1}{T^2}\sum_t Var(Y_t|U)\right] + Var\left[\frac{1}{T}\sum_t E(Y_t|U)\right] =$$

$$= \frac{1}{T^2}\sum_t v + Var(m(U)) = \frac{1}{T}v + w.$$

Inoltre, riesce

$$\operatorname{cov}(Y_{T+1}, \bar{Y}_T) = E\left[\operatorname{cov}(Y_{T+1}, \bar{Y}_T)|U\right] + \operatorname{cov}\left(E\left(Y_{T+1}|U\right), E\left(\bar{Y}_T|U\right)\right) =$$
$$= \operatorname{cov}\left(E\left(Y_{T+1}|U\right), E\left(\bar{Y}_T|U\right)\right) = \operatorname{cov}\left(m(U), m(U)\right) = Var(m(U)) = w.$$

Quindi si ha

$$b = \frac{w}{\frac{1}{T}v + w} = \frac{Tw}{v + Tw}$$

e lo stimatore di credibilità lineare di Y_{T+1} è

$$\tilde{Y}_{T+1} = m\left(1 - \frac{Tw}{v+Tw}\right) + \frac{Tw}{v+Tw}\bar{Y}_T\,.$$

Rimangono da stimare i parametri m, v e w.

Supposto di avere osservazioni su k rischi analoghi per T periodi, gli stimatori non distorti proposti dallo stesso Bühlmann per m, v e w, rispettivamente, sono i seguenti:

- $\frac{1}{k}\sum_{i=1}^{k}\bar{Y}_i$ dove $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T}Y_{it}$
- $\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{T-1}\sum_{t=1}^{T}\left(Y_{it}-\bar{Y}_i\right)^2\right)$
- $\frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{Y}_i-\bar{\bar{Y}}\right)-\frac{1}{T}\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{T-1}\sum_{t=1}^{T}\left(Y_{it}-\bar{Y}_i\right)^2\right)$

dove Y_{it} indica il numero aleatorio relativo all'individuo i e al periodo t e le condizioni di analogia tra i rischi sono le seguenti

- $Y_1\ Y_2 \quad Y$ i.i.d. condizionatamente a U_i
- $(U_1, Y_{11},...,Y_{1T}),...,(U_k, Y_{k1},...,Y_{kT})$ i.i.d.

Si osservi che, nella credibilità lineare empirica, per trovare i coefficienti dello stimatore di credibilità vengono utilizzati direttamente gli stimatori sopra indicati e dunque non occorre stimare la distribuzione del parametro di rischio.

Modello di Bühlmann-Straub

Passiamo ora ad un modello più generale in cui si suppone che $Y_t|U=u$, $t=1,2,...$ siano indipendenti, con uguale speranza matematica $E(Y_t|U)=m(U)$, ma non identicamente distribuiti.

Si suppone inoltre che

$$Var\left[Y_t|U=u\right]=\frac{1}{P_t}u^2(U).$$

I $P_1,...,P_T$ vengono detti *volumi* o *esposizioni*. Da queste ipotesi segue

$$E(Y_t) = E\left[E(Y_t|U)\right] = E\left[m(U)\right] = m$$

che non dipende da *t* e

$$Var(Y_t) = E\left[Var(Y_t|U)\right] + Var\left[E(Y_t|U)\right] = E\left[\frac{1}{P_t}u^2(U)\right] + Var\left[m(U)\right] = \frac{1}{P_t}v + w$$

con $v = E\left[u^2(U)\right]$, e $w = Var\left[m(U)\right]$.

Risolvendo il sistema (2.4.1) con le ipotesi fatte, si determinano $b_0, b_1, ..., b_T$. Si ottiene

$$b_0 = m\left(1 - \sum_t b_t\right)$$

$$b_t = \frac{P_t w}{v + Pw}, \quad t = 1, 2, ..., T;$$

dove $P = \sum_{t=1}^{T} P_t$.

Posto $k = \frac{v}{w}$, lo stimatore di credibilità di Y_{T+1} è

$$\tilde{Y}_{T+1} = m\frac{k}{k+P} + \frac{P}{k+P}\bar{\bar{Y}}_T,$$

dove $\bar{\bar{Y}}_T = \sum_t \frac{P_t}{P} Y_t$.

Rimangono ora da stimare i parametri m, v e w.

Con riferimento a *k* rischi analoghi, osservati per T periodi, supponiamo di disporre, per ciascun rischio, della coppia di osservazioni (Y_{it}, P_{it}). Posto $Z_{it} = \frac{Y_{it}}{P_{it}}$ si suppone:

- condizionatamente a U_i, $Z_{i1}, ..., Z_{iT}$ stocasticamente indipendenti per ogni $i = 1, ..., k$;
- $E\left(Z_{it}|U_i\right) = m\left(U_i\right)$ per ogni t, con $m(\cdot)$ indipendente da i;

- $Var\left(Z_{it}\middle|U_i\right)=\dfrac{1}{P_{it}}u^2\left(U_i\right)$ per ogni *t*, con $u(\cdot)$ indipendente da i;
- $\left(U_1,Z_{11},...,Z_{1T}\right),...,\left(U_k,Z_{k1},...,Z_{kT}\right)$ stocasticamente indipendenti;
- $U_1,...,U_k$ identicamente distribuiti.

Posto $P_{\cdot\cdot}=\sum_i P_{i\cdot}=\sum_i\sum_t P_{it}$, $P^*=\dfrac{1}{kT-1}\sum_{i=1}^{h}P_{i\cdot}\left(1-\dfrac{P_{i\cdot}}{P_{\cdot\cdot}}\right)$, vengono proposti come stimatori di m, v e w, rispettivamente, i seguenti numeri aleatori:

$\dfrac{1}{P_{\cdot\cdot}}\sum_{i=1}^{k}P_{i\cdot}\bar{\bar{Z}}_i$, dove $\bar{\bar{Z}}_i=\dfrac{1}{P_{i\cdot}}\sum_{t=1}^{T}P_{it}Z_{it}$;

$\dfrac{1}{k}\sum_{i=1}^{k}\left(\dfrac{1}{T-1}\sum_{t=1}^{T}P_{it}\left(Z_{it}-\bar{\bar{Z}}_i\right)^2\right)$;

$\dfrac{1}{P^*}\left[\dfrac{1}{kT-1}\sum_{i=1}^{k}\sum_{t=1}^{T}P_{it}\left(Z_{it}-\bar{\bar{Z}}\right)^2-\dfrac{1}{k}\sum_{i=1}^{k}\left(\dfrac{1}{T-1}\sum_{t=1}^{T}P_{it}\left(Z_{it}-\bar{\bar{Z}}_i\right)^2\right)\right]$,

con $\bar{\bar{Z}}=\dfrac{1}{P_{\cdot\cdot}}\sum_i P_{i\cdot}\bar{\bar{Z}}_i$. Si può provare che i tre stimatori proposti sono non distorti.

2.5 STIMATORI DI CREDIBILITÀ PER I MODELLI CON ETEROGENEITÀ

Vediamo ora una applicazione dello stimatore di credibilità di Bühlmann-Straub al modello con eterogeneità per il numero di sinistri presentato al Paragrafo 1.3.

Sia N_t il numero di sinistri per un fissato individuo nel periodo *t*. Supponiamo $N_t\middle|U=u\sim P(l_t u)$, dove u tiene conto della rischiosità dell'individuo mentre l_t tiene conto della rischiosità della classe cui appartiene. Si supponga che $N_1\middle|U=u, N_2\middle|U=u,...$ siano indipendenti.

Sotto queste condizioni, posto $\widehat{N}_t = \frac{N_t}{l_t}$, $t = 1, 2, \ldots$, è immediato verificare che per il processo $\widehat{N}_1, \widehat{N}_2, \ldots$ valgono le ipotesi del modello di Bühlmann-Straub, e riesce

$$E\left(\widehat{N}_t \middle| U = u\right) = \frac{1}{l_t} E\left(N_t \middle| U = u\right) = \frac{1}{l_t} u\, l_t = u,$$

$$Var\left(\widehat{N}_t \middle| U = u\right) = \frac{1}{l_t^2} E\left(N_t \middle| U = u\right) = \frac{1}{l_t^2} u\, l_t = \frac{u}{l_t}.$$

Pertanto $m(U) = u$ e $u^2(U) = u$. Supponiamo ora $E(U) = 1$. Allora riesce

$$m = E\left[E\left(\widehat{N}_{it} \middle| U\right)\right] = E(U) = 1,$$

$$w = Var\left[E\left(\widehat{N}_{it} \middle| U\right)\right] = Var(U),$$

$$v = E\left[u^2(U)\right] = E(U) = 1$$

In queste ipotesi lo stimatore di credibilità lineare di Bühlmann-Straub per $\widehat{N}_{T+1}$ è dato da

$$\tilde{\widehat{N}}_{T+1} = \frac{k}{k+P} + \frac{P}{k+P} \bar{\bar{N}}_T,$$

dove $P = \sum_{t=1}^{T} l_t$, $k = \frac{v}{w} = w^{-1}$, ed inoltre si è posto $\bar{\bar{N}}_T = \sum_{t=1}^{T} \frac{l_t}{P} \widehat{N}_t$.

Concludiamo questo paragrafo presentando l'espressione dello stimatore di credibilità lineare per il premio puro riportato da J. Pinquet in alcuni dei suoi articoli (si veda per esempio Pinquet (2001b)).

Consideriamo il modello per il premio puro con eterogeneità descritto al Paragrafo 1.6 nel caso dei costi con distribuzione log-normale. Invece dell'intera distribuzione di $U = \begin{pmatrix} U_n \\ U_c \end{pmatrix}$, supponiamo di conoscere solo la sua speranza matematica e la matrice di varianze covarianze. Siano dunque

$E(U_n) = E(U_c) = 0$,

$Var(U_n) = V_{nn}$,

$Var(U_c) = V_{cc}$,

$\text{cov}(U_n, U_c) = V_{nc}$.

Dalle ipotesi del modello segue che per $Z_{T+1} = \sum_{j=1}^{N_{T+1}} C_{T+1j}$ riesce

$$E\left[Z_{T+1} \middle| U_n = u_n, U_c = u_c\right] = E\left[N_{T+1} \middle| U_n = u_n\right] E\left[C_{T+1j} \middle| N_{T+1} = n_{T+1}, U_c = u_c\right] =$$

$$= l_{T+1} e^{u_n} e^{m_{T+1} + u_c + \frac{1}{2}s^2} = l_{T+1} e^{m_{T+1} \frac{1}{2}s^2} e^{u_n} e^{u_c}.$$

Poniamo $W_n = e^{U_n}$, $W_c = e^{U_c}$. Il coefficiente bonus-malus relativo al premio è dunque

$$\frac{E\left[Z_{T+1} \middle| U_n, U_c\right]}{E\left[Z_{T+1}\right]} = \frac{l_{T+1} e^{m_{T+1} + \frac{1}{2}s^2} W_n W_c}{E\left[l_{T+1} e^{m_{T+1} + \frac{1}{2}s^2} W_n W_c \right]} = \frac{W_n W_c}{E\left(W_n W_c\right)}.$$

L'obiettivo, in un approccio di credibilità lineare, è quello di fornire uno stimatore di questo coefficiente tramite una funzione del tipo

$$1 + a_n (N - l^*) + a_c (TLC - tlc^*), \qquad (2.5.1)$$

dove si è posto $N = \sum_{t=1}^{T} N_t$, $l^* = E(N)$, $TLC = \sum_{t,j} \log C_{tj}$, $tlc^* = E(\sum_{t,j} \log C_{tj})$.

In altre parole si vuole di risolvere il seguente problema di ottimo

$$\min_{a_n, a_c} E_q \left[\left(\frac{W_n W_c}{E(W_n W_c)} - [1 + a_n (N - l^*) + a_c (TLC - tlc^*)] \right)^2 \right].$$

Indicata con $Q(a_n, a_c)$ la funzione da ottimizzare, si ha

[2.5]

$$Q(a_n, a_c) = E_q\{\left(\frac{W_n W_c}{E(W_n W_c)}\right)^2 + 1 + a_n^2 (N - l^*)^2 + a_c^2 (TLC - tlc^*)^2 -$$
$$-2\frac{W_n W_c}{E(W_n W_c)}\left(1 + a_n (N - l^*) + a_c (TLC - tlc^*)\right) + 2a_n (N - l^*) +$$
$$+2a_c (TLC - tlc^*) + 2a_n a_c (N - l^*)(TLC - tlc^*)\} =$$
$$= E_q\left(\frac{W_n W_c}{E(W_n W_c)}\right)^2 + 1 + a_n^2 E_q (N - l^*)^2 + a_c^2 E_q (TLC - tlc^*)^2 -$$
$$-2E_q\left(\frac{W_n W_c}{E(W_n W_c)}\right) - 2a_n E_q\left(\frac{W_n W_c}{E(W_n W_c)}(N - l^*)\right) - 2a_c E_q\left(\frac{W_n W_c}{E(W_n W_c)}(TLC - tlc^*)\right) +$$
$$+2a_n a_c E_q\left((N - l^*)(TLC - tlc^*)\right).$$

Considerando il sistema delle condizioni del primo ordine,

$$\begin{cases} \dfrac{\partial Q}{\partial a_n} = 0 \\ \dfrac{\partial Q}{\partial a_c} = 0 \end{cases}$$

si ottiene che i coefficienti a_n e a_c sono le soluzioni del sistema lineare

$$\begin{cases} m_{nn} a_n + m_{nc} a_c = b_n \\ m_{cn} a_n + m_{cc} a_c = b_c \end{cases}$$

con

$$m_{nn} = E_q\left[(N - l^*)^2\right];$$
$$m_{nc} = E_q\left[(N - l^*)(TLC - tlc^*)\right];$$
$$m_{cn} = m_{nc};$$
$$m_{cc} = E_q\left[(TLC - tlc^*)^2\right];$$
$$b_n = E_q\left[(N - l^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right];$$
$$b_c = E_q\left[(TLC - tlc^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right].$$

[2.5]

Ovvero con semplici calcoli riesce:

$$a_n = \frac{m_{cc} b_n - m_{nc} b_c}{m_{nn} m_{cc} - (m_{nc})^2};$$

$$a_c = \frac{m_{nn} b_c - m_{nc} b_n}{m_{nn} m_{cc} - (m_{nc})^2}.$$

Se sono dati i valori dei momenti precedentemente elencati, sostituendo la soluzione del sistema nella (2.5.1), si ottiene lo stimatore cercato.

Allora uno stimatore di credibilità del premio individuale $E\left[Z_{T+1} \middle| U_n, U_c\right]$ è

$$\left[1 + a_n (N - l^*) + a_c (TLC - tlc^*)\right] E(Z_{T+1}).$$

Nel prossimo capitolo vengono fornite le espressioni degli stimatori proposti da Pinquet per i parametri del modello.

CAPITOLO 3

STIME PER I MODELLI CON ETEROGENEITÀ

3.1 INTRODUZIONE

L'applicazione dei modelli con eterogeneità per la tariffazione in base all'esperienza richiede di stimare i parametri del modello. Tali parametri sono quelli delle distribuzioni condizionate ai diversi parametri di rischio e quelli che individuano la distribuzione misturante o alcuni suoi momenti.

In questo capitolo descriviamo alcuni metodi che vengono utilizzati per la stima dei modelli con eterogeneità.

Il primo criterio che illustriamo è quello della massima verosimiglianza richiamando alcune importanti proprietà degli stimatori.

Come si è detto nel capitolo precedente, per rendere più flessibili i modelli, talvolta si può scegliere di non assegnare una specifica distribuzione misturante, ma di fornire solo alcuni dei suoi momenti. I modelli di questo tipo si dicono modelli mistura semi-parametrici, e un metodo che viene spesso utilizzato in questi casi, e di cui diamo un cenno in questo capitolo, è il metodo dei momenti.

Il principale problema che si incontra nella stima dei parametri dei modelli con eterogeneità risiede nel fatto che la verosimiglianza può avere una espressione analitica difficile da trattare e quindi la stima dei parametri può riuscire particolarmente complessa.

Per ovviare a questo inconveniente, nei suoi articoli Pinquet propone un procedimento di stima che in qualche modo combina i due metodi sopra citati. In

questo capitolo presentiamo tale procedimento facendo riferimento in particolare all'articolo "Linear credibility predictors for the pure premium of an insurance contract" (2001b).

In sostanza, il metodo di stima proposto si articola in due fasi successive. Nella prima fase si calcolano le stime di massima verosimiglianza del modello base. Viene quindi calcolato il cosiddetto valore pseudo-vero che è il limite in probabilità dello stimatore del modello base. La seconda fase prevede la stima di alcuni momenti della distribuzione misturante attraverso i residui calcolati nel modello base.

Il capitolo si conclude con la presentazione degli stimatori dei modelli con eterogeneità illustrati nel Capitolo 1.

3.2 MODELLI BASATI SULLA VEROSIMIGLIANZA.

In questo paragrafo richiamiamo il procedimento di stima della massima verosimiglianza ricordando alcune proprietà dei corrispondenti stimatori. Per la stesura ci siamo basati su Cameron, Trivedi (1998).

Consideriamo una sequenza di variabili aleatorie scalari $Y_1,...,Y_K$ stocasticamente indipendenti e supponiamo che dati il vettore delle componenti di regressione x_k e il vettore di parametri $q_1 \in \Theta_1$, Y_k sia distribuita con densità $\ell(y_k \,|\, x_k, q_1)$ $k = 1,...,K$. Indichiamo con q_1 il vettore dei parametri perché, come si vedrà nel seguito, le stime di massima verosimiglianza vengono effettuate per il modello base. Il principio della massima verosimiglianza indica di scegliere come stima di q_1 il valore appartenente a Θ_1 che massimizza la funzione di probabilità o di densità congiunta di $(Y_1,...,Y_K)$, calcolata nel campione osservato $y_1,..., y_k$. Tale probabilità o densità, vista come una funzione dei parametri viene detta funzione di verosimiglianza, e per le ipotesi accolte si ha

$$L(q_1) = \prod_{k=1}^{K} \ell(y_k \,|\, x_k, q_1) .$$

Si osservi che massimizzare la funzione di verosimiglianza equivale a massimizzare la funzione di log-verosimiglianza

$$l(q_1) = \log L(q_1) = \sum_{k=1}^{k} \log \ell(y_k \,|x_k, q_1)\,.$$

Se sono soddisfatte opportune condizioni di regolarità (si veda ad esempio Cameron, Trivedi pag. 23), la stima di massima verosimiglianza $\hat{q_1}$, è la soluzione delle condizioni del primo ordine:

$$\frac{\partial l}{\partial q_1} = \sum_{k=1}^{K} \frac{\partial \log \ell(y_k \,|x_k, q_1)}{\partial q_1} = 0\,,$$

dove $\frac{\partial l}{\partial q_1}$ è un vettore $q \times 1$, se q_1 è un vettore con q componenti. Il vettore $\frac{\partial l}{\partial q_1}$ viene anche detto vettore *score*. Si dice poi stimatore di massima verosimiglianza del parametro q_1, basato su $(Y_1,...,Y_K)$, il numero aleatorio funzione di $(Y_1,...,Y_K)$ che associa ad ogni k-upla $(y_1,...,y_K)$, determinazione possibile del vettore $(Y_1,...,Y_K)$, la stima di massima verosimiglianza di q_1 basata su $(y_1,...,y_K)$. Indichiamo con $\tilde{q}_1 = \hat{q_1}(Y_1,...,Y_K)$ lo stimatore di massima verosimiglianza basato su $(Y_1,...,Y_K)$.

Osserviamo che al variare di K otteniamo una successione di stimatori.

I risultati classici di consistenza e normalità asintotica degli stimatori di massima verosimiglianza valgono se sono soddisfatte opportune condizioni di regolarità. Ricordiamo che lo stimatore $\tilde{q}_1$ è detto *consistente* se

$$\lim_{K \to +\infty} \Pr_{q_1}\left(\left|\tilde{q}_1 - q_1\right| \geq e\right) = 0\,, \text{ per ogni } e > 0\,^{5},$$

[5] Con i simboli $\Pr_{q_1}$, E_{q_1}, Var_{q_1} indichiamo in questo paragrafo la probabilità, la speranza matematica e la varianza calcolate con la distribuzione di $(Y_1,...,Y_K)$ di parametro q_1.

ovvero se la successione (dipendente da K) $\tilde{q}_1$ converge in probabilità a q_1, quando si assume per $(Y_1,...,Y_K)$ la distribuzione di parametro q_1.

Per quanto riguarda la distribuzione asintotica dello stimatore di massima verosimiglianza si ha

$$\sqrt{K}(\tilde{q}_1 - q_1) \xrightarrow{d} N(0, A^{-1}),\ K \to +\infty,$$

dove la matrice $q \times q$, A, è definita come segue:

$$A = -\lim_{K\to\infty} \frac{1}{K} E\left[\sum_{k=1}^{K} \frac{\partial^2 \log \ell(y_k | x_k, q_1)}{\partial q_1 \partial q_1'} \bigg|_{q_1} \right].$$

Quando si assume che la distribuzione del vettore aleatorio $(Y_1,...,Y_K)$ abbia parametro q_1 appartenente ad un insieme Θ_1 e si determina la stima di massima verosimiglianza di q_1 in Θ_1 si parla di approccio della massima verosimiglianza con distribuzione "completamente specificata". Vediamo ora che cosa si intende per approccio della massima verosimiglianza in un contesto di non specificazione.

Sia $(\ell_{q_1})_{q_1 \in \Theta_1}$ una famiglia parametrica di funzioni di densità e supponiamo che le densità della famiglia attribuiscano probabilità nulla agli stessi boreliani di $\mathbb{R}$, si dice anche in questo caso che le distribuzioni sono "equivalenti". Poniamo $\tilde{q}_1^0 = \arg q_1 \max_{q_1 \in \Theta_1} \sum_{k=1}^{K} \log \ell_{q_1}(Y_k)$ [6], dove $(Y_k)_{k=1,...,K}$ è un vettore di variabili i.i.d. con una distribuzione Q di densità ℓ_Q.

[6] Data una funzione $f(x)$ definita in un insieme D a valori reali, dotata di massimo assoluto, indichiamo con il simbolo $\arg x \max_{x \in D} f(x)$ il punto di massimo assoluto. Analogamente con $\arg x \min_{x \in D} f(x)$ indichiamo il minimo assoluto.

Se la distribuzione ℓ_Q non appartiene a $(\ell_{q_1})_{q_1\in\Theta_1}$ allora si dice che il modello è mal specificato. Si noti che la ℓ_Q indica una generica densità, non necessariamente una della famiglia cui appartengono le ℓ_{q_1}.

Supposto che la distribuzione ℓ_Q sia equivalente a quelle della famiglia $(\ell_{q_1})_{q_1\in\Theta_1}$ poniamo

$$E_Q[f(Y)] = \int f(y)\ell_Q(y)dy$$

la speranza matematica di $f(Y)$ se Y ha distribuzione di densità ℓ_Q. Analogamente poniamo

$$E_{q_1}[f(Y)] = \int f(y)\ell_{q_1}(y)dy.$$

In questo ambito sussiste il seguente risultato (si veda Pinquet (2001a) per i riferimenti):

$$Q\lim_{K\to+\infty}\tilde{q}_1^0 = \arg q_1 \max_{q_1\in\Theta_1} E_Q(\log\ell_{q_1}(y)) = \arg q_1 \min_{q_1\in\Theta_1} KL(Q/\ell_{q_1}), \qquad (3.2.1)$$

dove $KL(Q/\ell_{q_1}) = E_Q(\log\ell_Q(y) - \log\ell_{q_1}(y))$ è un indice di dissimilarità tra misure di probabilità equivalenti detto di Kullback-Leibler e dove il simbolo $Q\lim_{K\to+\infty}\tilde{q}_1^0$ indica il limite in probabilità rispetto alla distribuzione di densità ℓ_Q, della successione degli stimatori $\tilde{q}_1^0$ al divergere di K. Il limite di $\tilde{q}_1^0$, $q_1^*(Q)$, viene detto *valore pseudo-vero.*

Osserviamo che, dalla prima uguaglianza nella (3.2.1) in condizioni che consentono la derivazione sotto il segno di integrale, segue che il valore pseudo vero è tale che

$$E_Q\left[\left.\frac{\partial\log\ell_{q_1}(y)}{\partial q_1}\right|_{q_1=q_{1(Q)}^*}\right] = 0 \qquad (3.2.2)$$

è quindi tale che la speranza matematica dello score calcolata in $q_1^*(Q)$, rispetto alla distribuzione ℓ_Q è nulla.

In una impostazione di tipo classico, si assume che il vettore aleatorio $(Y_1,...,Y_K)$ di cui si dispone delle osservazioni $(y_1,...,y_K)$ abbia una distribuzione ben definita che però non è nota.

Le osservazioni vengono utilizzate per cercare di individuare o almeno approssimare tale distribuzione "vera". In quest'ottica se la distribuzione "vera" non appartiene alla famiglia $(\ell_{q_1})_{q_1\in\Theta_1}$, il valore pseudo-vero è un valore per il parametro che rappresenta la soluzione meno sfavorevole con riferimento all'indice di dissimilarità di Kullback-Leibler. La figura 3.2.1 illustra questo aspetto.

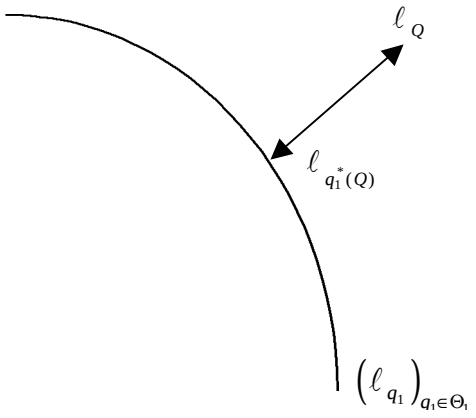


Fig. 3.2.1. Valore Pseudo-Vero

- *Esempio 1*

Consideriamo una famiglia di distribuzioni equivalenti parametrizzate dalla speranza matematica: $(\ell_m)_{m\in M}$ e supponiamo che le densità abbiano una struttura esponenziale lineare, cioè che riesca

$$\ell_m(y) = \exp[A(m) + B(y) + C(m)y].$$

Allora si prova che la distribuzione della famiglia $(\ell_m)_{m\in M}$ che minimizza l'indice di Kullback-Leibler rispetto ad una distribuzione di densità ℓ_Q equivalente a quella della famiglia è quella che ha speranza matematica

$$m^*(Q) = \int y\ell_Q(y)\partial y\,,$$

(per riferimenti si veda Pinquet (2001a)). Quindi se invece di considerare la distribuzione ℓ_Q si considera una distribuzione della famiglia $(\ell_m)_{m\in M}$, la soluzione meno sfavorevole è quella con parametro $m = m^*(Q)$.

- *Esempio 2*

Consideriamo ora un caso in cui le distribuzioni dipendono dalle componenti di regressione. Supponiamo di disporre di una sequenza di osservazioni. I dati siano relativi ad un processo di conta, ad esempio i numeri dei sinistri che colpiscono un contratto. Inoltre, si supponga che il processo che genera i dati sia un processo di numeri con distribuzione mistura di distribuzioni di Poisson.

Le componenti di regressione *x* siano collegate con il parametro q_1, la varianza della distribuzione misturante sia s^2.

Indichiamo con N la variabile di cui si vuole valutare la distribuzione supponiamo che si abbia

$$E(N\,|q_1, s^2, x) = \exp(x'q_1)\,.$$

Pertanto il valore atteso di N non dipende dalla varianza della distribuzione misturante. In queste condizioni si dice che il modello con eterogeneità è ben specificato rispetto alla speranza matematica. Questo è il caso del modello binomiale-negativo.

Consideriamo ora la famiglia delle distribuzioni di Poisson con le stesse componenti di regressione. Le distribuzioni di Poisson hanno una struttura esponenziale lineare, perciò per quanto visto nell'Esempio 1, la distribuzione di

Poisson con il parametro pari a $m = \exp(x'q_1)$ minimizza l'indice di dissimilarità di Kullback-Leibler rispetto alla distribuzione di N. Dal momento che questo risultato vale qualsiasi sia il valore di x, vettore delle componenti di regressione, per la (3.1.1) si ottiene che il limite in probabilità di $\tilde{q}_1^0$, ovvero dello stimatore di massima verosimiglianza per il modello di Poisson è q_1:

$$\left(q_1, s^2\right) \lim_{K \to +\infty} \tilde{q}_1^0 = q_1 .$$

Dunque si conclude che lo stimatore di massima verosimiglianza per il modello di Poisson fornisce stime consistenti per i parametri del modello con eterogeneità a lui collegato, qualora quest'ultimo sia ben specificato con riferimento al valore atteso.

3.3 METODI BASATI SUI MOMENTI

I metodi che si basano sui momenti forniscono stimatori in ambito semi-parametrico.

Si supponga che $(Y_k)_{k=1,\ldots,K}$ sia una sequenza di variabili i.i.d. con valori in $\mathbb{R}^m$, e la distribuzione dipenda da un vettore q di parametri e da un vettore di componenti di regressione *x*. Si consideri una famiglia parametrizzata di funzioni $(f_q)_{q \in \Theta}$, con $f_q : \mathbb{R}^m \to \mathbb{R}^s$. Le funzioni f_q hanno componenti legate ai momenti della distribuzione degli $Y_1,\ldots,Y_K$ e possono dipendere anche dalle componenti di regressione.

Lo stimatore $\hat{q}^{GMM}$ di q del metodo dei momenti generalizzato (si veda Pinquet(2001a)) si ottiene risolvendo rispetto a q la:

$$\frac{1}{K} \sum_{k=1}^{K} f(Y_k) = 0 .$$

Sotto opportune ipotesi lo stimatore è consistente.

Applichiamo questo procedimento di stima al modello di Poisson con eterogeneità ben specificato con riferimento alla speranza matematica dell'Esempio 2, Paragrafo 3.2. Si ha:

$$Var(N|q_1, s^2, x) = E\left[Var\left(N|U\right)\right] + Var\left[E\left(N|U\right)\right] =$$

$$= E(N^2|q_1, s^2, x) + s^2 E^2(N|q_1, s^2, x),$$

con $E(N|q_1, s^2, x) = \exp(x'q_1)$.

Poniamo

$$q = \begin{pmatrix} q_1 \\ s^2 \end{pmatrix} \text{ e } f_q(n) = \begin{pmatrix} (n - \exp(x'q_1))x \\ (n - \exp(x'q_1))^2 - n - s^2[\exp(x'q_1)]^2 \end{pmatrix}.$$

Sia $n_1,...,n_K, x_1,...,x_K$ un campione rispettivamente di valori osservati e di componenti di regressione. Allora le stime del metodo dei momenti generalizzato per il modello di Poisson con eterogeneità sono:

$$\hat{q_1} = \sum_k \left(n_k - \hat{l}_k\right) x_k \,, \qquad \hat{s}^2 = \frac{\sum_k \left[(n_k - \hat{l}_k)^2 - n_k\right]}{\sum_k \hat{l}_k^2}.$$

Osserviamo che $\hat{q_1}$ coincide con la stima di massima verosimiglianza per il modello di Poisson, $\hat{q_1}^0$, definita dalla relazione ortogonale $\sum_k (n_k - \hat{l}_k)x_k = 0$, $\hat{l}_k = \exp(x_k \hat{q_1}^0)$. Si noti che lo stimatore di s^2 non è vincolato ad essere positivo; un segno positivo significa sovra-dispersione dei residui.

3.4 IL METODO DI STIMA PROPOSTO DA PINQUET

Come abbiamo già detto, nei modelli con eterogeneità la verosimiglianza può avere un'espressione che rende difficile una trattazione analitica della stessa

ai fini di ottenere le stime dei parametri. Illustriamo allora il procedimento di stima proposto da J. Pinquet (si veda ad esempio Pinquet (2001a)).

In sintesi risulta che stimatori consistenti per i modelli con eterogeneità descritti nel Capitolo 1 si possono ottenere attraverso:

- Il calcolo di un valore pseudo-vero ottenuto come limite dello stimatore di massima verosimiglianza del modello base.
- La stima di alcuni momenti della distribuzione mistura attraverso i residui calcolati nel modello base.

Presentiamo il metodo a partire da un esempio.

Si consideri il modello di Poisson con eterogeneità. Sia $N_1, N_2, \ldots$ il processo di interesse e U il parametro di rischio. I numeri aleatori del processo siano stocasticamente indipendenti condizionatamente a U. Indicato per brevità con N il generico numero aleatorio del processo si ha

$$N \mid U = u \sim P(\exp(x'q_1)u)$$

dove x è il vettore delle componenti di regressione. Si noti che il vettore x è in generale diverso per i diversi numeri aleatori della successione. Allora

$$E(N) = \exp(x'q_1)E(U) = \exp(x'(q_1 + \log[E(U)]e_1)),$$

dove si è supposto che l'intercetta sia la prima delle k componenti di regressione, ovvero che $x' = \left(1, x_1, \ldots x_{k_1-1}\right)$ e che e_1 sia il primo vettore della base canonica di $\mathbb{R}^k$, $e_1 = (1,0,\ldots,0)$. Osserviamo che la speranza matematica di N nel modello con eterogeneità coincide con la speranza matematica del modello di Poisson di parametro $q_1^*(q_1, s^2) = q_1 + \log[E(U)]e_1$ se le componenti di regressione sono le stesse.

Nel Paragrafo 3.2 abbiamo ricordato che la distribuzione di Poisson con parametro pari a $m = \exp(x'(q_1 + \log[E(U)]e_1)$ minimizza l'indice di dissimilarità di Kullback-Leibler rispetto alla distribuzione mistura di distribuzioni di Poisson.

Poiché questo risultato vale qualunque sia x, dalla (3.2.1) segue che il valore pseudo-vero è pari a $q_1^* = q_1 + \log[E(U)]e_1$. Ciò significa che q_1^* è il limite in probabilità di $\tilde{q}_1^0$, stimatore di massima verosimiglianza del modello di Poisson, quando si assume che il processo che genera i dati sia mistura di distribuzioni di Poisson. Si ottiene di conseguenza (si veda S.C. Port. (1994) Proposizione 49.5):

$$\tilde{E}^0(N) = \exp(x'\tilde{q}_1^0) \to \exp(x'q_1^*) = \exp(x'q_1)E(U) = E(N)\,. \qquad (3.4.1)$$

dove il simbolo $\tilde{E}^0(N)$ indica lo stimatore della speranza matematica di *N* calcolato ipotizzando il modello base e la convergenza è in probabilità. Si ha quindi che $\tilde{E}^0(N)$ è uno stimatore consistente di $E(N)$.

Grazie all'interpretazione data nel Paragrafo 1.2 per un modello con random effects questo risultato può venire espresso nel seguente modo: il numero atteso di sinistri di un individuo, calcolato nel modello a priori, converge al numero atteso di sinistri dell'individuo generico a lui collegato, se il processo che genera i dati ha come distribuzione quella del modello con random effects. Tale proprietà vale qualunque siano i valori dei fattori tariffari e qualunque sia la distribuzione misturante.

La varianza di *N* nel modello con random effects è data da:

$$\begin{aligned} &Var(N) = E\left[Var(N|U)\right] + Var\left[E(N|U)\right] = E\left(\exp(x'q_1)U\right) + Var\left(\exp(x'q_1)U\right) = \\ &= E(N) + [\exp(xq_1)^2]Var(U) \end{aligned} \qquad (3.4.2)$$

Vediamo ora come a partire dallo stimatore di massima verosimiglianza $\tilde{q}_1^0$ del modello di Poisson, possiamo ottenere uno stimatore consistente per uno dei momenti della distribuzione misturante. Precisamente, determiniamo uno stimatore per il quadrato del coefficiente di variazione di *U*. Indicato con $CV(U) = \dfrac{s(U)}{E(U)}$ il coefficiente di variazione di *U*, si ha

$$CV^2(U) = \frac{Var(U)}{E^2(U)} = \frac{Var(N) - E(N)}{E^2(N)}$$

dove l'ultima uguaglianza deriva dalla (3.4.2) ricordando che $E(N) = \exp(xq_1)E(U)$. Da quest'ultima relazione si ricava:

$$CV^2(U) = E\left[\frac{(N - E(N))^2 - N}{E^2(N)}\right].$$

Ciò suggerisce di considerare come stimatore di $CV(U)^2$ il numero aleatorio

$$CV^2(U) = \frac{\sum_{k=1}^{K}\left[\left(N_k - \tilde{E}^0(N_k)\right)^2 - N_k\right]}{\sum_{k=1}^{K}\tilde{E}^{0^2}(N_k)}.$$

Poiché, con passaggi analoghi a quelli che hanno portato alla (3.4.2), ricordando le proprietà del processo $N_1|U = u, N_2|U = u, \ldots$ si ha

$$Var\left(\sum_{k=1}^{K} N_k\right) = E\left(\sum_{k=1}^{K} N_k\right) + E^2\left(\sum_{k=1}^{K} N_k\right)CV^2(U),$$

ovvero

$$E\left(\sum_{k=1}^{K} N_k - E\left(\sum_{k=1}^{K} N_k\right)\right) = E\left(\sum_{k=1}^{K} N_k\right) + E^2\left(\sum_{k=1}^{K} N_k\right)CV^2(U),$$

dal limite (3.4.1), segue che tale stimatore è consistente.

Pertanto, supposto di disporre dei valori osservati di $N_1, \ldots, N_K$ e delle rispettive componenti di regressione, posto $\hat{l}_k = \exp\left(x_k'\hat{q}_1^0\right)$, con $\hat{q}_1^0$ stima di massima verosimiglianza del modello di Poisson, si può porre

$$\widehat{CV(U)^2} = \frac{\sum_k (n_k - \hat{l}_k)^2 - n_k}{\sum_k \hat{l}_k^2}. \qquad (3.4.3)$$

Dunque il quadrato del coefficiente di variazione della distribuzione misturante viene stimato in maniera consistente a partire dallo stimatore di

massima verosimiglianza del modello di Poisson. Dalle equazioni (3.4.1), (3.4.2), (3.4.3) si ottiene una stima per $Var(N)$ nel modello con eterogeneità:

$$\widehat{Var(N)} = \hat{l}_k + (\hat{l}_k^2 CV^2(U)), \text{ per ogni } k.$$

In questo modo si è ottenuta una stima consistente per la varianza di *N*. Questi risultati sono di grande utilità per il calcolo dei previsori di credibilità lineare.

L'esempio precedente può venire generalizzato anche per gli altri modelli con eterogeneità del Capitolo 1. Basandoci su Pinquet (2001b) descriviamo il procedimento per il modello relativo al premio puro quando per i risarcimenti si consideri la distribuzione log-normale.

Ricordiamo che il modello prevede che il processo di rischio per l'*i*-esimo assicurato $i = 1,.., p$, sia $Y_i = (N_{it}, C_{it1}, C_{it2}, ...)_{t=1,...,T_i}$. Nel modello con eterogeneità il parametro di rischio è bidimensionale $U_i = \begin{pmatrix} U_{ni} \\ U_{ci} \end{pmatrix}$. Le ipotesi probabilistiche sono precisate nel Paragrafo 1.6.

Indichiamo con $S_{q_1,x_i}(y_i)$ lo *score* calcolato per l'i-esimo assicurato nel modello tariffario a priori, dove $x_i = (x_{it})_{t=1,...,T_i}$ è la sequenza delle componenti di regressione e dove $y_i = (n_{it}, c_{itj})_{t=1,...,T_i;1\le j\le n_{it}}$ è la sequenza delle osservazioni relative alle variabili aleatorie che descrivono il rischio.

Dalla (1.6.1) otteniamo:

$$S_{q_1,x_i} = \begin{pmatrix} \frac{\partial l}{\partial b} \\ \frac{\partial l}{\partial a} \\ \frac{\partial l}{\partial s^2} \end{pmatrix} = \begin{pmatrix} \sum_t (n_{it} - l_{it}) x_{it} \\ \sum_{t,j} \frac{1}{s^2} (\log c_{itj} - m_{it}) x_{it} \\ \frac{1}{2s^4} \sum_{t,j} \left[(\log c_{itj} - m_{it})^2 - s^2 \right] \end{pmatrix} \qquad (3.4.4)$$

dove $l_{it} = \exp(x_{it}' b)$, $m_{it} = x_{it}' a$.

La stima $\hat{q}_1^0$ di massima verosimiglianza di $q_1 = \begin{pmatrix} b \\ a \\ s^2 \end{pmatrix}$ è la soluzione della seguente equazione di verosimiglianza nell'incognita q_1

$$\overline{S_{q_1}} = \frac{1}{p}\sum_{i=1}^{p} S_{q_1,x_i}(y_i) = 0\,. \tag{3.4.5}$$

Si supponga ora che il processo che genera i dati appartenga al modello con eterogeneità. Indichiamo al solito con $q = (q_1, q_2)$ il vettore dei parametri del modello e con $E_{q,x_i}(N_{it})$, $E_{q,x_i}\left(\sum_{j=1}^{N_{it}} \log C_{itj}\right)$ le speranze matematiche di N_{it} e della somma dei logaritmi dei risarcimenti per l'individuo *i*, all'epoca *t*, calcolate nel modello con eterogeneità. Ricordiamo che $q_2 = \begin{pmatrix} V_{nn} \\ V_{nc} \\ V_{cc} \end{pmatrix}$.

Usiamo il simbolo $E_{q,x}(\cdot)$ per indicare la speranza matematica di un numero aleatorio funzione delle variabili di rischio dei diversi assicurati del portafoglio.

Per questo modello sussiste la seguente proprietà:

esiste una funzione $q_1^*(\cdot)$ tale che $E_{q,x}\left(S_{q_1^*(q),x}\right) = 0$, per ogni *x* e per ogni q. (3.4.6)

Si può provare che il valore $q_1^*(q)$, che non dipende dalle distribuzione componenti di regressione, risulta essere il valore pseudo-vero, ovvero il valore a cui converge, in probabilità, lo stimatore di massima verosimiglianza del modello a priori:

$$(q,x)\lim \tilde{q}_1^0 = \begin{pmatrix} \tilde{a}^0 \\ \tilde{b}^0 \\ \widetilde{s^2}^0 \end{pmatrix} = q_1^*(q) = \begin{pmatrix} a^*(q) \\ b^*(q) \\ s_*^2(q) \end{pmatrix}, \text{ per ogni } q \text{ e per ogni } x. \tag{3.4.7}$$

Osserviamo che $q_1^*(q)$ è tale che la speranza matematica dello score calcolato in $q_1^*(q)$, valutata rispetto alla distribuzione mistura di parametro q e componenti di regressione *x*, sia nulla per ogni q e per ogni *x* (si veda (3.2.2)).

Verifichiamo l'esistenza dei valori pseudo-veri $a^*(q)$, $b^*(q)$ e $s_*^2(q)$. Precisamente, proviamo che esistono $a^*(q)$, $b^*(q)$ e $s_*^2(q)$ tali che

$$E_{q,x}\left(S_{q_1^*(q),x}\right)=0, \text{ per ogni } i,\ x_i \text{ e } q.$$

Consideriamo separatamente i tre blocchi dello score (3.4.4). La condizione diventa

$$\sum_t\left[E_{q,x_i}\left(N_{it}-\exp(x_{it}'b^*(q)\right)x_{it}\right]=0,$$

$$\frac{1}{s^2}\sum_t E_{q,x_i}\left[\sum_{j=1}^{N_{it}}\left(\log C_{itj}-x_{it}'a^*(q)\right)x_{it}\right]=0,$$

$$\frac{1}{2s^4}\sum_t E_{q,x_i}\left[\sum_{j=1}^{N_{it}}\left(\log C_{itj}-x_{it}'a^*(q)\right)^2-s_*^2(q)N_{it}\right]=0.$$

Proviamo che esistono $a^*(q)$, $b^*(q)$ e $s_*^2(q)$ tali che

$$E_{q,x_i}\left(N_{it}-\exp(x_{it}'b^*(q)\right)=0,$$

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}}\left(\log C_{itj}-x_{it}'a^*(q)\right)\right]=0,$$

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}}\left(\log C_{itj}-x_{it}'a^*(q)\right)^2-s_*^2(q)N_{it}\right]=0,$$

per ogni q, x_i, e *t*. Da questo segue ovviamente che sono soddisfatte le tre precedenti condizioni e quindi la $E_{q,x}\left(S_{q_1^*(q),x}\right)=0$. Le equazioni sopra scritte sono equivalenti alle

$$E_{q,x_i}(N_{it})=\exp(x_{it}'b^*(q)),$$

[3.4]

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}}(\log(C_{itj})-x_{it}'a^*(q))\right]=0,$$

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}}(\log(C_{itj})-x_{it}'a^*(q))^2\right]=E_{q,x_i}(N_{it})s_*^2(q).$$

Poiché abbiamo supposto che la distribuzione di (U_{ni},U_{ci}) sia invariante al variare di i poniamo, per semplicità di scrittura, $(U_{ni},U_{ci})=(U_n,U_c)$ e, $W_n=\exp(U_n)$, $W_c=\exp(U_c)$. Si ha allora

$$E_{q,x_i}(N_{it})=\exp(x_{it}'b)E(W_n)=\exp(x_{it}'(b+\log[E(W_n)]e_1)=\exp(x_{it}'b^*(q)), \quad (3.4.8)$$

se l'intercetta è la prima delle k componenti di regressione nel modello di Poisson e se e_1 è il primo vettore di base canonica $\mathbb{R}^k$. Poiché U_n ha distribuzione normale di speranza matematica nulla e varianza V_{nn}, si ha

$$E(W_n)=E[\exp(U_n)]=\exp(V_{nn}/2),$$

dove l'ultima uguaglianza si ottiene osservando che $E[\exp(U_n)]$ è il valore in $t=1$ della funzione generatrice dei momenti della distribuzione di U_n, $m_{U_n}(t)=E\left[\exp\left(tU_n\right)\right]=\exp\left(\frac{V_{nn}t}{2}\right)$, (si veda la nota a piè di pagina 4). Pertanto, riesce

$$b^*(q)=b+\log[E(W_n)]e_1=b+\frac{V_{nn}}{2}e_1.$$

Calcoliamo ora $a^*(q)$.

Si deve risolvere l'equazione in a^*:

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}}(\log(C_{itj})-x_{it}'a^*)\right]=0. \quad (3.4.9)$$

Poiché

[3.4]

$$E\left[\sum_{j=1}^{N_{it}}(\log(C_{itj})-x_{it}'a^{*})\middle|U_n=u_n,U_c=u_c\right]=$$

$$=E\left[N_{it}\middle|U_n=u_n\right]E\left[(\log(C_{itj})-x_{it}'a^{*})\middle|N_{it}=n_{it},U_c=u_c\right]=$$

$$=\exp\left(x_{it}'b\right)\exp\left(u_n\right)\left(x_{it}'a+u_c-x_{it}'a^{*}\right),$$

dalla disintegrabilità della speranza matematica si ha

$$E_{q,x_i}\left\{E\left[\sum_{j=1}^{N_{it}}(\log(C_{itj})-x_{it}'a^{*})\middle|U_n,U_c\right]\right\}=$$

$$=E_{q,x_i}\left\{\exp\left(x_{it}'b\right)\exp\left(U_n\right)\left(x_{it}'a+U_c-x_{it}'a^{*}\right)\right\}=$$

$$=l_{it}E\left(W_n\right)\left[x_{it}'\left(a-a^{*}\right)+\frac{E\left(W_nU_c\right)}{E\left(W_n\right)}\right].$$

L'equazione (3.4.9) diventa

$$l_{it}E\left(W_n\right)\left[x_{it}'\left(a-a^{*}\right)+\frac{E\left(W_nU_c\right)}{E\left(W_n\right)}\right]=0,$$

se l'intercetta è la prima delle k componenti di regressione nel modello e se e_1 è il primo vettore di base canonica $\mathbb{R}^k$. Il valore pseudo vero è

$$a^{*}(q)=a+\frac{E(W_nU_c)}{E(W_n)}e_1=a+V_{cn}e_1, \qquad (3.4.10)$$

dove l'ultima uguaglianza segue dalla $E(W_nU_c)=V_{cn}E(W_n)$.

Per ricavare quest'ultima osserviamo intanto che se X è un numero aleatorio con distribuzione $N(0,1)$, allora

$$E\left[\exp(j\,X)X^{k}\right]=\exp\left(\frac{j^{2}}{2}\right)E\left[(X+j)^{k}\right] \text{ per ogni } j\in\mathbb{R} \text{ e per ogni } k\in\mathbb{N}, \qquad (3.4.11)$$

come facilmente si verifica.

Ora, dalla formula di scomposizione di Choleski della matrice di varianze-covarianze del parametro aleatorio bidimensionale $\begin{pmatrix} U_n \\ U_c \end{pmatrix}$, (si veda il Paragrafo 2.3 Esempio 5), si ha

$$E(W_n U_c) = E\left[\exp(U_n) U_c\right] = E\left[\exp\left(j_{nn} S_n\right)\left(j_{cn} S_n + j_{cc} S_c\right)\right] =$$
$$= j_{cn} E\left[\exp\left(j_{nn} S_n\right) S_n\right] + j_{cc} E\left[\exp\left(j_{nn} S_n\right) S_c\right],$$

dove S_n, S_c sono stocasticamente indipendenti con distribuzione $N(0,1)$. Allora sfruttando anche la (3.4.11) si ricava

$$E(W_n U_c) = j_{cn} \exp\left(\frac{j_{nn}^2}{2}\right) E\left[S_n + j_{nn}\right] = j_{cn} j_{nn} \exp\left(\frac{j_{nn}^2}{2}\right) = V_{cn} E\left(W_n\right).$$

Procedendo in modo analogo, dalla (3.4.11) e dalla formula di scomposizione di Cholesky si ricavano i seguenti risultati di interesse per i calcoli che seguiranno:

$$E(W_n W_c) = E\left(W_n\right) \exp\left[\frac{E\left(W_n U_c\right)}{E\left(W_n\right)} + \frac{V_{cc}}{2}\right];$$

$$E(W_n^2) = \exp(V_{nn}) E^2(W_n);$$

$$E(W_n^2 U_c) = 2V_{cn} \exp(V_{nn}) E^2(W_n);$$

$$E(W_n U_c^2) = (V_{cc} + V_{cn}^2) E(W_n);$$

$$E(W_n^2 U_c^2) = \left[\exp(V_{nn})(V_{cc} + 4V_{cn}^2)\right] E^2(W_n);$$

$$E(W_n^2 W_c) = \exp(V_{nn} + V_{cn}) E\left(W_n\right) E(W_n W_c);$$

$$E(W_n^2 W_n U_c) = (V_{cc} + 2V_{cn}) \exp(V_{nn} + V_{cn}) E(W_n) E(W_n W_c). \qquad (3.4.12)$$

L'ultimo valore pseudo-vero $s_*^2(q)$ è la soluzione dell'equazione in s_*^2

$$E_{q,x_i}\left[\sum_{j=1}^{N_{it}} (\log(C_{itj}) - x_{it}' a^*(q))^2\right] = E_{q,x_i}(N_{it}) s_*^2. \qquad (3.4.13)$$

Osserviamo che il numero aleatorio $\log(C_{itj}) - x'_{it}a^*(q) \,|\, N_{it} = n_{it}, U_c = u_c$ ha distribuzione normale di varianza s^2 e speranza matematica $x'_{it}a + u_c - x'_{it}a^*(q)$. Dalla (3.4.10), si ha $x'_{it}\left(a - a^*(q)\right) + u_c = x'_{it}\left(-V_{cn}e_1\right) + u_c = -V_{cn} + u_c$, dove l'ultima uguaglianza deriva dall'ipotesi che l'intercetta sia la prima delle componenti di x'_{it} e e_1 è il primo vettore della base canonica di $\mathbb{R}^k$. Il momento secondo, somma della varianza e del quadrato della speranza matematica, è dunque $s^2 + (u_c - V_{cn})^2$.

Allora, per la proprietà di disintegrabilità, la (3.4.13) può venire riscritta come segue

$$E_{q,x}\left\{\exp(x'_{it}b)\exp(U_n)\left[s^2 + (U_c - V_{cn})^2\right]\right\} = E_{q,x}\left[\exp(x'_{it}b)\exp(U_n)\right]s_*^2.$$

Si ha allora l'equazione

$$\exp(x'_{it}b)\,E_{q,x_i}(W_n[(U_c - V_{cn})^2 + s^2 - s_*^2]) = 0,$$

equivalente alla

$$E(W_nU_c^2) + V_{cn}^2E(W_n) - 2V_{cn}E(W_nU_c) + \left(s^2 + s_*^2\right)E(W_n) = 0.$$

Dalle $E(W_nU_c) = V_{cn}E\left(W_n\right)$ e $E(W_nU_c^2) = (V_{cc} + V_{cn}^2)E(W_n)$ (si veda 3.4.12), si ottiene:

$$s_*^2(q) = s^2 + V_{cc}. \tag{3.4.14}$$

Considerando ora per l'assicurato i-esimo i seguenti numeri aleatori,

$$\sum_{t=1}^{T_i} N_{it}\,,\ \sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}} \log(C_{itj})\,,\ \sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}} C_{itj}\,, \tag{3.4.15}$$

e gli stimatori di massima verosimiglianza di $\tilde{a}^0$, $\tilde{b}^0$, $\widetilde{s^2}^0$ del modello a priori.

[3.4]

Se indichiamo con $\tilde{E}^0(\cdot)$ lo stimatore della speranza matematica nel modello a priori si ha

$$\tilde{E}^0\left(\sum_{t=1}^{T_i} N_{it}\right)=\sum_{t=1}^{T_i}\exp(x_{it}'\tilde{b}^0),$$

$$\tilde{E}^0\left(\sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}}\log(C_{itj})\right)=\sum_{t=1}^{T_i}\exp(x_{it}'\tilde{b}^0)(x_{it}'\tilde{a}^0),$$

$$\tilde{E}^0\left(\sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}} C_{itj}\right)=\sum_{t=1}^{T_i}\exp(x_{it}'\tilde{b}^0)\exp(x_{it}'\tilde{a}^0+\frac{\widetilde{s^2}^0}{2}).$$

Poiché per la (3.4.7) $\tilde{a}^0$, $\tilde{b}^0$, $\widetilde{s^2}^0$ convergono in probabilità (nella distribuzione del modello con eterogeneità) a $a^*(q)$, $b^*(q)$ e $s_*^2(q)$, allora i tre precedenti stimatori convergono in probabilità rispettivamente a

$$\sum_{t=1}^{T_i}\exp(x_{it}'b^*(q))=\sum_{t=1}^{T_i}\exp(x_{it}'b)E(W_n)=\sum_{t=1}^{T_i}E_{q,x_i}(N_{it}),$$

$$\sum_{t=1}^{T_i}\exp(x_{it}'b^*(q))(x_{it}'a^*(q))=\sum_{t=1}^{T_i}\exp(x_{it}'b)E(W_n)\left(x_{it}'a+\frac{E(W_nU_c)}{E(W_n)}\right)=$$

$$=\sum_{t=1}^{T_i}E_{q,x_i}\left(\sum_{j=1}^{N_{it}}\log C_{itj}\right)=E_{q,x_i}\left(\sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}}\log C_{itj}\right),$$

$$\sum_{t=1}^{T_i}\exp(x_{it}'b^*(q))\exp\left(x_{it}'a^*(q)+\frac{s_*^2(q)}{2}\right)=$$

$$=\sum_{t=1}^{T_i}\exp(x_{it}'b)E(W_n)\left(x_{it}'a+\frac{E(W_nU_c)}{E(W_n)}+\frac{s^2+V_{cc}}{2}\right)=$$

$$=\sum_{t=1}^{T_i}E_{q,x_i}\left(\sum_{j=1}^{N_{it}}C_{itj}\right)=E_{q,x_i}\left(\sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}}C_{itj}\right),$$

per la penultima uguaglianza si sfrutta la prima delle (3.4.12).

Se indichiamo con $\bar{Y}_i$ il vettore di componenti (3.4.15), possiamo sintetizzare il precedente risultato con la

$$\tilde{E}^0\left(\bar{Y}_i\right) \to E_{q,x_i}\left(\bar{Y}_i\right), \text{ per ogni } i, \tag{3.4.16}$$

dove la convergenza è in probabilità.

Tale proprietà consente di ottenere stimatori consistenti per alcuni momenti della distribuzione mistura attraverso i residui calcolati nel modello tariffario a priori. Per i modelli di Poisson con random effects, i previsori di credibilità lineare si possono calcolare a partire dalla stima di questi momenti.

A titolo di esempio nel prossimo paragrafo mostreremo come dai precedenti risultati relativi al modello per il premio puro si ottengono stimatori consistenti dei parametri che compaiono nel previsore di credibilità lineare nel Paragrafo 2.5.

3.5 PREVISORI DI CREDIBILITÀ LINEARE E STIMATORI CONSISTENTI

Con riferimento all'*i*-esimo assicurato, per semplicità di scrittura, poniamo

$$N_i = \sum_{t=1}^{T_i} N_{it}\,,\ \ l_i^* = E_{q,x_i}(N_i)\,,\ \ \tilde{l}_i = \tilde{E}^0(N_i)\,,\ \ \hat{l}_i = \hat{E}^0(N_i)\,,$$

$$TLC_i = \sum_{t=1}^{T_i}\sum_{j=1}^{N_{it}} \log C_{itj}\,,\ \ tlc_i^* = E_{q,x_i}(TLC_i)\,,\ \ \widetilde{tlc}_i = \tilde{E}^0(TLC_i)\,,\ \ \widehat{tlc}_i = \hat{E}^0(TLC_i)\,,$$

dove come in precedenza, il simbolo $\tilde{E}^0(\cdot)$ indica uno stimatore della speranza matematica del numero aleatorio, ottenuto utilizzando gli stimatori di massima verosimiglianza dei parametri nel modello a priori e il simbolo $\hat{E}^0(\cdot)$ indica la corrispondente stima.

Come illustrato nel Paragrafo 2.5, si assuma per l'*i*-esimo assicurato il seguente coefficiente bonus-malus:

$$1 + a_{ni}(n_i - \hat{l}_i) + a_{ci}(tlc_i - \hat{E}^0(TLC_i))\,. \tag{3.5.1}$$

Dal limite sintetizzato dalla (3.4.16) si ha

- $\tilde{l}_i = \tilde{E}^0(N_i) = \sum_{t=1}^{T_i} \exp(x_{it}' \tilde{b}^0)$ converge in probabilità a $l_i^* = l_i E(W_n)$,

  con $l_i = \sum_{t=1}^{T_i} \exp(x_{it}' b)$.

- $\tilde{E}^0(TLC_i) = \sum_{t=1}^{T_i} \exp(x_{it}' \tilde{b}^0)(x_{it}' \tilde{a}^0)$ converge in probabilità a tlc_i^*.

Nel Paragrafo 2.5, abbiamo mostrato che i coefficienti a_{ni} e a_{ci}, scritti ora per l'i-esimo assicurato, sono

$$a_{ni} = \frac{m_{cc}^i b_{ni} - m_{nc}^i b_{ci}}{m_{nn}^i m_{cc}^i - (m_{nc}^i)^2}$$

$$a_{ci} = \frac{m_{nn}^i b_{ci} - m_{nc}^i b_{ni}}{m_{nn}^i m_{cc}^i - (m_{nc}^i)^2} \qquad (3.5.2)$$

con

$$m_{nn}^i = E_{q,x_i}\left[(N_i - l_i^*)^2\right];$$

$$m_{nc}^i = E_{q,x_i}\left[(N_i - l_i^*)(TLC_i - tlc_i^*)\right];$$

$$m_{cc}^i = E_{q,x_i}\left[(TLC_i - tlc_i^*)^2\right];$$

$$b_{ni} = E_{q,x_i}\left[(N_i - l_i^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right];$$

$$b_{ci} = E_{q,x_i}\left[(TLC_i - tlc_i^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right].$$

Calcoleremo simultaneamente le stime di questi momenti e dei parametri della distribuzione mistura, supponendo di disporre delle osservazioni dei processi $(N_{it}, C_{it1}, C_{it2}, \ldots)$, $t = 1, \ldots, T_i$, $i = 1, \ldots, p$.

Poiché $E_{q,x_i}(N_i) = l_i^*$, si ha $E_{q,x_i}\left[(N_i - l_i^*)^2\right] = Var_{q,x_i}(N_i)$. D'altra parte, riesce (si vedano le osservazioni seguenti alla (3.4.2)) $Var_{q,x_i}(N_i) = E_{q,x_i}(N_i) + E_{q,x_i}^2(N_i)CV^2(W_n)$ e $\tilde{l}_i$ converge a $l_i^* = E_{q,x_i}(N_i)$, allora si ha che:

$$\hat{V}_{nn}^1 = \frac{\sum_i \left[(n_i - \hat{l}_i)^2 - n_i\right]}{\sum_i \hat{l}_i^2};$$

è una stima consistente (valore osservato di uno stimatore consistente) di $CV^2(W_n)$ dove $n_i = \sum_{t=1}^{T_i} n_{it}$. Allora si ha

$$\hat{Var}_{q,x_i}(N_i) = \hat{l}_i + \hat{l}_i^2 \hat{V}_{nn}^1.$$

Dal momento che $CV^2(W_n) = \frac{Var(W_n)}{E^2(W_n)} = \frac{E(W_n^2) - E^2(W_n)}{E^2(W_n)} = \frac{E(W_n^2)}{E^2(W_n)} - 1 =$ $\exp(V_{nn}) - 1$, dove l'ultima uguaglianza segue dalla seconda delle (3.4.12), otteniamo

$$\hat{V}_{nn} = \log(1 + \hat{V}_{nn}^1),$$

che è una stima consistente di V_{nn}.

Poiché si è supposto che, subordinatamente al numero di sinistri, i costi siano indipendenti, nel fixed effects model si ottiene:

$$E\left[(N_i - l_i^*)(TLC_i - tlc_i^*) \middle| U_{ni} = u_{ni}, U_{ci} = u_{ci}\right] = l_i(w_{ni} - E(W_n))(l_i w_{ni}(u_{ci} - V_{cn})).$$

Per ottenere la precedente uguaglianza si passa attraverso la proprietà di disintegrabilità della speranza matematica condizionata rispetto alla partizione del generico evento $\left(N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i} \middle| U_{ni} = u_{ni}, U_{ci} = u_{ci}\right)$, dove $\left(n_{i1}, ..., n_{iT_i}\right)$ indica una determinazione possibile di $\left(N_{i1}, ..., N_{iT_i}\right)$.

La speranza matematica nel random effects model è pari a

[3.5]

$$E_{q,x_i}\left[(N_i - l_i^*)(TLC_i - tlc_i^*)\right] = l_i^2 E\left[(W_n - E(W_n)(W_n(U_c - V_{cn}))\right] =$$

$$= l_i^2 E^2(W_n)[V_{cn} + 2V_{cn}\exp(V_{nn}) - V_{cn} - V_{cn}\exp(V_{nn})] = l_i^2 E^2(W_n)[V_{cn}\exp(V_{nn})],$$

dove si sono utilizzate le uguaglianze (3.4.12).

Sfruttando i due limiti richiamati all'inizio di questo paragrafo, dalla precedente uguaglianza otteniamo che

$$\hat{V}_{cn} = \frac{\sum_i \left(n_i - \hat{l}_i\right)\left(tlc_i - \widehat{tlc_i}\right)}{\sum_i \hat{l}_i^2\left(1+\hat{V}_{nn}^1\right)},$$

dove $tlc_i = \sum_{t=1}^{T_i} \log c_{itj}$, è uno stimatore consistente di V_{cn}. Dunque si ottiene la seguente stima

$$\hat{E}_{q,x_i}\left[(N_i - l_i^*)(TLC_i - tlc_i^*)\right] = \hat{l}_i^2\hat{V}_{cn}\exp(\hat{V}_{nn}) = \hat{l}_i^2\hat{V}_{cn}(1+\hat{V}_{nn}^1).$$

Sfruttando le ipotesi del modello e le (3.4.12), si prova che

$$E_{q,x_i}\left[(TLC_i - tlc_i^*)^2\right] = E\left[\, l_i W_n + l_i^2 W_n^2)(U_c - V_{cn})^2 + l_i W_n s^{\,2}\right] =$$

$$= l_i^*(V_{cc} + s^{\,2}) + (l_i^*)^2\exp(V_{nn})(V_{cn}^2 + V_{cc}) = l_i^* s_*^{\,2} + (l_i^*)^2\exp(V_{nn})(V_{cn}^2 + V_{cc}),$$

dove s_*^2 indica il valore pseudo-vero dato dalla (3.4.14). Il limite $\widetilde{s^{\,2}}^0 \to s_*^2$ conduce allo stimatore consistente di V_{cc}. Si ottiene dunque

$$\hat{V}_{cc} = \frac{\sum_i\left[\left(tlc_i - \widehat{tlc_i}\right) - n_i\widehat{s^{\,2}}^0\right]}{\left(\sum_i \hat{l}_i^2\right)\left(1+\hat{V}_{nn}^1\right)} - \hat{V}_{cn}^2,$$

e quindi

$$\hat{E}_{q,x_i}\left[(TLC_i - tlc_i^*)^2\right] = \hat{l}_i\hat{s}^{\,2^0} + \hat{l}_i^2\left[(1+\hat{V}_{nn}^1)(\hat{V}_{cn} + \hat{V}_{cc})\right]$$

[3.5]

Dobbiamo ancora ottenere delle stime per le espressioni del secondo membro del sistema lineare:

$$b_{ni} = E_{q,x_i}\left[(N_i - l_i^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right] \text{ e } b_{ci} = E_{q,x_i}\left[(TLC_i - tlc_i^*)\left(\frac{W_n W_c}{E(W_n W_c)}\right)\right].$$

Nel fixed effects model si ha

$$E\left[(N_i - l_i^*)\left(\frac{w_{ni} w_{ci}}{E(W_n W_c)}\right)\middle| U_{ni} = u_{ni}, U_{ci} = u_{ci}\right] = l_i\left[(w_{ni} - E(W_n)\left(\frac{w_{ni} w_{ci}}{E(W_n W_c)}\right)\right].$$

Allora la speranza matematica b_{ni} nel random effects model è pari a

$$l_i E(W_n)\left[\frac{E(W_n^2 W_c)}{E(W_n)E(W_n W_c)} - 1\right] = l_i^*\left[\frac{E(W_n^2 W_c)}{E(W_n)E(W_n W_c)} - 1\right].$$

Attraverso le uguaglianze fornite nella (3.4.12), una stima consistente è:

$$b_{ni} = \hat{l}_i\left[\exp(\hat{V}_{nn} + \hat{V}_{cn}) - 1\right].$$

Analogamente, nel fixed effects model riesce:

$$E\left[(TLC_i - tlc_i^*)\left(\frac{w_{ni} w_{ci}}{E(W_n W_c)}\right)\middle| U_{ni} = u_{ni}, U_{ci} = u_{ci}\right] = l_i w_{ni}(u_{ci} - V_{cn})\left(\frac{w_{ni} w_{ci}}{E(W_n W_c)}\right).$$

La speranza matematica b_{ci} nel random effects model è allora pari a

$$l_i E(W_n)\left[\frac{E(W_n^2 W_c U_c)}{E(W_n)E(W_n W_c)} - V_{cn}\frac{E(W_n^2 W_c)}{E(W_n)E(W_n W_c)}\right],$$

ed è stimata in maniera consistente da

$$b_{ci} = \hat{l}_i(\hat{V}_{cn} + \hat{V}_{cc})\exp(\hat{V}_{nn} + \hat{V}_{cn}).$$

Le stime dei momenti così ottenute, sostituite nelle (3.5.2) e poi nella (3.5.1) conducono alla formula di credibilità lineare per il coefficiente bonus-malus.

3.6 STIMATORI PER GLI ALTRI MODELLI CON ETEROGENEITÀ

In questo paragrafo riportiamo gli stimatori per i modelli con eterogeneità citati nel Capitolo1.

In Pinquet (2001a) è dichiarato che gli stimatori che vengono forniti sono espliciti, consistenti, asintoticamente normali e asintoticamente efficienti nell'ipotesi nulla. Si ricordi a tale proposito che l'ipotesi nulla è collegata al modello base ovvero all'assenza di eterogeneità osservata.

a. *Modello ad una sola equazione per il numero di sinistri, con una funzione di autocorrelazione costante per il random effect.*

Il modello è descritto nel Paragrafo 1.3, Esempio 3. Ricordiamo che

$$N_{it}\,|U_{it}=u_{it}\sim P(l_{it}u_{it});\qquad l_{it}=\exp(x_{it}'q_1)\qquad U_{it}=R_iS_{it},\qquad u_{it}=r_is_{it}.$$

Oltre a q_1, i parametri del modello con eterogeneità sono $s_r^{\,2}=Var(R_i)$ e $s_s^{\,2}=Var(S_{it})$. Se i valori attesi di R_i e S_{it} sono pari a 1, allora il modello con random effects è ben specificato con riferimento al valore atteso, e le stime consistenti per $s_r^{\,2}$ e $s_s^{\,2}$ sono:

$$\hat{s}_r^{\,2}=\hat{s}_r^{\,2^1};$$

$$\hat{s}_s^{\,2}=\frac{\hat{s}_s^{\,2^1}}{1+\hat{s}_r^{\,2^1}},$$

con $\hat{s}_r^{\,2^1}=\dfrac{\sum_i\sum_{t\neq t'}(n_{it}-\hat{l}_{it})(n_{it'}-\hat{l}_{it'})}{\sum_i\sum_{t\neq t'}\hat{l}_{it}\,\hat{l}_{it'}}$,

$$\hat{s}_r^{\,2^1}+\hat{s}_s^{\,2^1}=\frac{\sum_{i,t}\left[(n_{it}-\hat{l}_{it})^2-n_{it}\right]}{\sum_{i,t}\hat{l}_{it}^2}.$$

Gli stimatori $\hat{s}_r^{21}$ e $\hat{s}_s^{21}$ sono ottenuti attraverso una linearizzazione dello score calcolato per $q_1 = \hat{q}_1^{0}$ e $q_2 = 0$. Si considera cioè l'approssimazione di Taylor dello score (si veda Pinquet (1996)). Tali stimatori sono forniti in forma semiparametrica dal momento che sono stati ottenuti senza una specificazione parametrica della distribuzione mistura. Si osservi che

$\hat{s}_s^{21}, \hat{s}_s^{2} > 0$ se e solo se $\sum_i \sum_{t \neq t'} (n_{it} - \hat{l}_{it})(n_{it'} - \hat{l}_{it'}) > 0$;

$\hat{s}_r^{21}, \hat{s}_s^{21} > 0$ se e solo se $\dfrac{\sum_{i,t} \left[(n_{it} - \hat{l}_{it})^2 - n_{it} \right]}{\sum_{i,t} \hat{l}_{it}^2} > \dfrac{\sum_i \sum_{t \neq t'} (n_{it} - \hat{l}_{it})(n_{it'} - \hat{l}_{it'})}{\sum_i \sum_{t \neq t'} \hat{l}_{it} \hat{l}_{it'}}$

se e solo se $\dfrac{\sum_{i,t} \left[(n_{it} - \hat{l}_{it})^2 - n_{it} \right]}{\sum_{i,t} \hat{l}_{it}^2} > \dfrac{\sum_i \left[(n_i - \hat{l}_i)^2 - n_i \right]}{\sum_{i,t} \hat{l}_i^2}$. (3.6.1)

Gli stimatori $\hat{s}_r^{21}$ e $\hat{s}_s^{21}$ sono positivi se i residui di un individuo che sono calcolati in diversi periodi di tempo hanno lo stesso segno.

b. Il modello ad una sola equazione per il numero dei sinistri con funzione di autocorrelazione variante con il tempo per i random effects.

Il modello è descritto nel Paragrafo 1.3, Esempio 4. Partiamo dal modello con fixed effects.

$$N_{it} \left| U_{it} = u_{it} \sim P(l_{it} w_{it}); \; w_{it} = \exp(u_{it}) \right.$$

Il processo $(U_{it})_{t \geq 1}$ è stazionario e la distribuzione di U_{it} è normale di media nulla e varianza s^2 per ogni *i*. Poniamo

$s^2 = Var(U_{it})$;

$Cov(U_{it+h}, U_{it}) = s^2 r(h)$.

[3.6]

Non viene specificata la distribuzione del processo $(U_{it})_{t\geq 1}$ ma si fornisce uno stimatore consistente per s^2 e per la funzione di autocorrelazione.

Le stime sono

$$\hat{s}^2 = \log\left(\frac{\sum_{i,t}\left[(n_{it} - \hat{l}_{it})^2 - n_{it}\right]}{\sum_{i,t}\hat{l}_{it}^2}\right),$$

$$\hat{s}^2\hat{r}(h) = \log\left[1 + \frac{\sum_{i|T_i>h}\sum_{T_i\geq t>h}(n_{it} - \hat{l}_{it})(n_{it-h} - \hat{l}_{it-h})}{\sum_{i|T_i>h}\sum_{T_i\geq t>h}\hat{l}_{it}\hat{l}_{it-h}}\right]$$

Osserviamo che queste stime sono ottenute senza porre vincoli ai momenti della distribuzione misturante, pertanto i $\hat{r}(h)$ possono non appartenere all'intervallo $[-1,1]$.

c. Modello a più equazioni per il numero di sinistri.

Con le notazioni utilizzate al Paragrafo 1.4 le stime degli elementi della matrice varianze-covarianze del parametro vettoriale di rischio V_{jj}, V_{jl} $(j \neq l)$ sono

$$\hat{V}_{jj}^1 = \frac{\sum_i\left[\left(n_i^{(j)} - \hat{l}_i^{(j)}\right)^2 - \hat{l}_i^{(j)}\right]}{\sum_i \hat{l}_i^{(j)2}};$$

$$\hat{V}_{jl}^1 = \frac{\sum_i\left(n_i^{(j)} - \hat{l}_i^{(j)}\right)\left(n_i^{(l)} - \hat{l}_i^{(l)}\right)}{\sum_i \hat{l}_i^{(j)}l_i^{(l)}} \quad (j \neq l).$$

Come già in precedenza, anche in questo caso gli stimatori di V_{jj} e V_{jl} sono ricavati dopo aver effettuato una linearizzazione dello score.

Se $W_j = \exp(U_j)$, dove U_j ha la distribuzione di U_{ij}, j-esima componente del parametro di rischio dell'assicurato i-esimo si può dimostrare che

$$\tilde{V}_{jl}^{1} \to \frac{E[W_j W_l]}{E[W_j]E[W_l]} \text{ per ogni } j,l\,. \qquad (3.6.2)$$

Questa proprietà conduce ai previsori di credibilità lineare ottenuti tramite un approccio semiparametrico.

In un approccio parametrico, si suppone $U_i \sim N_q(0,V)$, dove questo simbolo indica la distribuzione normale di media nulla e matrice varianze-covarianze V per un vettore *q*-dimensionale.

Poiché $\frac{E[W_j W_l]}{E[W_j]E[W_l]} = \exp(V_{jl}) - 1$, dalla (3.6.2) si ottiene che

$$\hat{V}_{jl} = \log(1+\hat{V}_{jl}^{1}), \text{ per ogni } j,l\,,$$

fornisce uno stimatore consistente di V.

CAPITOLO 4

UN'APPLICAZIONE NUMERICA AI MODELLI CON ETEROGENEITÀ

4.1 INTRODUZIONE.

Abbiamo considerato un portafoglio di 44885 polizze di assicurati RCA osservate in tre anni consecutivi per complessive 134655 osservazioni. Gli anni in questione sono il 1998, il 1999 e il 2000.

Per ciascuna polizza si dispone di 11 informazioni, e precisamente:

- il totale, espresso in Euro, degli indennizzi dei sinistri con seguito, espresso dalla somma tra importo pagato e riservato;
- il numero dei sinistri con seguito riportati dall'assicurato;
- la durata del periodo di osservazione dell'assicurato misurata in anni, dove si è considerato l'anno composto da 360 giorni;
- l'anno di inizio dell'osservazione della polizza, variabile introdotta per indurre un effetto generazionale.

Le seguenti informazioni su variabili che descrivono l'individuo:

- sesso;
- età;
- provincia di residenza;
- professione;

e infine le informazioni che riguardano le caratteristiche del veicolo:

- i cavalli fiscali;
- la potenza misurata in kilowatt;
- l'età del veicolo.

Visto l'elevato numero di determinazioni di alcune variabili, attraverso la *cluster analysis*, abbiamo ottenuto una ripartizione in classi in modo da avere una suddivisione meno fine di quella iniziale. Classificazioni di questo tipo sono molto frequenti nella pratica e sono motivate con esigenze commerciali.

Per alcune variabili (potenza in kilowatt, cavalli fiscali, età ed anno di immatricolazione) si hanno determinazioni con esposizioni molto diverse tra loro, ed in particolare alcune molto elevate, altre piuttosto basse. Il metodo di aggregazione delle determinazioni può produrre classi poco significative se non procediamo preventivamente nell'aggregare "in modo manuale" alcune determinazioni così da ottenere, per ciascuna variabile, una ripartizione in classi della stessa con esposizioni abbastanza vicine.

Il metodo che abbiamo utilizzato per la cluster analysis è il *metodo Ward o della minima varianza interna* con due variabili risposta: la frequenza sinistri e la variabile stessa. Le variabili di interesse sono preventivamente standardizzate, in modo da eliminare le differenze di scala.

Si osservi che il costo medio non viene preso in considerazione come variabile risposta, perché è un dato stimato (costo liquidato+costo risarcito) inoltre, quand'anche fosse un dato "reale", si potrebbe verificare una riapertura della pratica con conseguente modifica dello stesso.

Richiamiamo il metodo Ward.

Data una variabile tariffaria con N determinazioni, siano $g_1,...,g_N$ le esposizioni complessive per ciascuna delle N determinazioni; g_h è precisamente l'esposizione complessiva delle polizze con determinazione *h* della variabile in oggetto.

Fissato l'*i*-esimo individuo, e la *h*-esima determinazione della variabile tariffaria, indicati con n_{ih} e g_{ih} rispettivamente il numero dei sinistri e l'esposizione, la frequenza sinistri è data da: $f_{ih} = \frac{n_{ih}}{g_{ih}}$.

Indicato con m_h il numero dei rischi con determinazione h si può calcolare la frequenza media per la h-esima determinazione:

$$f_h = \sum_{i=1}^{m_h} f_{ih} \frac{g_{ih}}{g_h},$$

e la varianza interna al gruppo con determinazione h:

$$s_h^2 = \sum_{i=1}^{m_h} \frac{g_{ih}}{g_h} (f_{ih} - f_h)^2.$$

Inoltre, posto $g = \sum_{h=1}^{N} g_h$, la varianza totale interna è data da

$${}_N s^2 = \sum_{h=1}^{N} s_h^2 \frac{g_h}{g}.$$

Il generico passo del procedimento è il seguente:

supposto di avere una ripartizione in K classi, con varianza totale interna ${}_K s^2 = \sum_{h=1}^{K} {}_K s_h^2 \frac{{}_K g_h}{g}$, ${}_K g_h$ essendo l'esposizione nella *h*-esima classe della ripartizione, si passa a K-1 classi scegliendo l'aggregazione di due classi che rende minima la varianza totale interna ${}_{K-1} s^2$, con ${}_{K-1} s^2 = \sum_{h=1}^{K-1} \frac{{}_{K-1} g_h}{g} {}_{K-1} s_h^2$.

Con due variabili risposta A e B, dopo averle standardizzate, supponendo l'indipendenza tra di esse, si considera la varianza totale interna ${}_K s_A^2 + {}_K s_B^2$ e si passa da K a K-1 classi raggruppando le due classi che danno il minimo di ${}_{K-1} s_A^2 + {}_{K-1} s_B^2$.

[4.1]

Per l'arresto del procedimento ci si basa sulla funzione R^2 definita come $1-\frac{{}_K s^2}{{}_1 s^2}$. Nell'esempio numerico il procedimento si arresta quando il coefficiente di correlazione R^2 riesce non inferiore a 0.95, ossia quando la perdita di informazione conseguente all'aver ripartito le determinazioni in classi è inferiore alla soglia del 5% che abbiamo fissato a priori. In alternativa ci si può arrestare quando si è raggiunto un numero prefissato di classi.

Per applicare il metodo di Ward ai nostri dati abbiamo utilizzato le procedure *cluster* e *tree* di SAS. Nella prossima fig. 4.1.1 riportiamo un esempio di output del programma SAS.

Cluster History

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ[7]	Tie
104	EE	RO	291	0.0000	1.00	T
103	CL104	RS	297	0.0000	1.00	
102	AN	BL	1104	0.0000	1.00	
101	BI	PN	1986	0.0000	1.00	
100	AV	LC	1248	0.0000	1.00	
99	AT	TE	843	0.0000	1.00	
98	KR	PD	3141	0.0000	1.00	
97	LO	TN	2034	0.0000	1.00	
96	GE	TV	5046	0.0000	1.00	
95	EN	GR	447	0.0000	1.00	
94	SA	VC	726	0.0000	1.00	
93	CL101	BZ	3966	0.0000	1.00	
92	MT	RE	804	0.0000	1.00	
91	AR	RG	867	0.0000	1.00	
90	CN	OR	1368	0.0000	1.00	
89	CL98	RI	3570	0.0000	1.00	
88	CZ	PV	1740	0.0000	1.00	
87	CS	PC	975	0.0000	1.00	
86	BG	IM	2589	0.0000	1.00	
85	AL	CL89	4458	0.0000	1.00	
84	MN	SO	1356	0.0000	1.00	
83	GO	PZ	1785	0.0000	1.00	
82	LE	VV	2016	0.0000	1.00	
81	PS	VE	3687	0.0000	1.00	
80	CL84	PI	2277	0.0000	1.00	
79	BR	CL92	1977	0.0000	1.00	
78	CB	TR	804	0.0000	1.00	
77	CL97	UD	4416	0.0000	1.00	
76	RM	TP	12807	0.0000	1.00	
75	CL95	CL83	2232	0.0000	1.00	
74	AQ	CL87	1497	0.0000	1.00	
73	SS	VR	3348	0.0000	1.00	
72	MS	TS	4674	0.0000	1.00	
71	NU	SI	954	0.0000	1.00	
70	BS	SV	5019	0.0000	1.00	
69	CE	RA	1566	0.0000	1.00	
68	CL94	SR	1524	0.0000	1.00	
67	BN	CL77	4770	0.0000	1.00	
66	CL99	PO	2037	0.0000	1.00	

[7] L'indice di correlazione *R-square*, viene indicato nel programma SAS con il simbolo RSQ.

[4.1]

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ
65	FI	PA	3054	0.0000	1.00
64	CH	CL90	589	0.0000	1.00
63	MC	PR	1767	0.0000	1.00
62	CL67	FE	5715	0.0000	1.00
61	AP	ME	1398	0.0000	1.00
60	CL80	CL81	5964	0.0000	1.00
59	IS	VI	2256	0.0000	1.00
58	FR	LI	1770	0.0000	1.00
57	CL88	CL75	3972	0.0000	1.00
56	FG	VT	1722	0.0000	1.00
55	CL102	CL59	3360	0.0000	1.00
54	CL65	CL73	7302	0.0000	1.00
53	PE	SP	783	0.0000	1.00
52	CL70	CL71	5973	0.0000	1.00
51	CL85	CL64	6225	0.0000	1.00
50	CL	PT	2199	0.0000	1.00
49	CL60	VA	7914	0.0000	1.00
48	CL69	CR	2478	0.0000	1.00
47	CL61	MO	3252	0.0000	1.00
46	CL100	CL86	3837	0.0000	1.00
45	CL66	NA	5430	0.0000	1.00
44	CA	VB	2268	0.0000	1.00
43	CL93	RC	5091	0.0000	1.00
42	CL58	TO	5601	0.0000	1.00
41	CT	CL68	3174	0.0000	1.00
40	LU	CL76	13845	0.0000	1.00
39	CL79	NO	3576	0.0000	1.00
38	MI	PG	10338	0.0000	1.00
37	BA	CL63	6693	0.0000	1.00
36	CL55	CL82	5376	0.0000	1.00
35	CL40	TA	14475	0.0000	1.00
34	CO	LT	3252	0.0000	1.00
33	CL52	CL49	14076	0.0001	.999
32	CL56	RN	2835	0.0001	.999
31	CL46	CL44	6105	0.0001	.999
30	CL78	CL96	5850	0.0001	.999
29	CL74	CL48	3975	0.0001	.999
28	CL62	CL57	9687	0.0001	.999
27	CL47	CL54	10554	0.0001	.999
26	CL30	CL41	9024	0.0001	.999
25	CL51	CL43	11316	0.0001	.999
24	CL36	FO	6576	0.0001	.999
23	CL45	CL72	10104	0.0002	.998
22	CL28	CL39	13263	0.0002	.998
21	CL32	CL53	3618	0.0004	.998
20	CL27	CL38	20892	0.0004	.997
19	AG	CL103	657	0.0006	.997
18	CL23	BO	12153	0.0007	.996
17	CL25	CL42	16917	0.0007	.995
16	CL37	CL34	9945	0.0010	.994
15	CL29	CL21	7593	0.0011	.993
14	CL22	CL33	27150	0.0018	.991
13	CL20	CL31	26997	0.0023	.989
12	CL19	CL91	1524	0.0025	.987
11	CL16	CL50	12144	0.0026	.984
10	CL18	CL35	26628	0.0034	.981
9	CL17	CL24	23493	0.0038	.977
8	CL14	CL26	36174	0.0048	.972
7	AO	CL11	12246	0.0051	.967
6	CL12	CL15	9117	0.0179	.950
5	CL13	CL10	53625	0.0214	.928
4	CL6	CL9	32610	0.0391	.889
3	CL4	CL8	68784	0.0949	.794
2	CL7	CL5	65871	0.1189	.675
1	CL3	CL2	134655	0.6749	.000

Fig. 4.1.1 Output della *proc cluster* per la variabile provincia di residenza

La colonna NCL indica il numero di clusters, mentre nelle due colonne --Clusters Joined--- vengono riportati i nomi dei clusters che sono stati

aggregati. Le osservazioni sono individuate tramite il valore formattato della variabile identificativa, mentre i clusters con due o più osservazioni sono indicati con CL*n*, CL*n* essendo il cluster ottenuto alla riga *n*. Il numero di osservazioni presenti nel nuovo cluster è riportato nella colonna FREQ termine che indica *Frequency of New Cluster*. Nella colonna SPRSQ, *Semipartial R-Squared*, viene riportato il decremento della varianza risultante dall'unione dei due cluster presenti sulla riga, in proporzione della varianza totale. La colonna RSQ indica invece la proporzione di varianza spiegata dai clusters. Infine, la lettera T nella colonna TIE indica la presenza, in corrispondenza di quel cluster, di un vincolo per la minima distanza; l'assenza di vincoli viene segnalata con uno spazio vuoto.

Nel caso della provincia si è utilizzata solo la frequenza sinistri come variabile risposta, e si è scelto di ripartire le determinazioni in sei macro province, dal momento che in corrispondenza del sesto cluster l'indice R^2 riesce pari a 0,95.

Il procedimento con due variabili risposta è stato usato con le seguenti variabili: età, cavalli fiscali, potenza in kilowatt ed immatricolazione.

Un discorso a parte deve essere fatto per la professione. Molte delle ventuno determinazioni di questa variabile sono di difficile definizione; per questo motivo si è scelto di accorpare a priori le determinazioni in modo da ottenere otto categorie professionali ben definite e abbastanza eterogenee e rappresentative della realtà.

A questo punto è stata effettuata un'analisi preliminare sulle variabili di interesse mettendo in evidenza per ogni classe di determinazioni la frequenza sinistri di classe e il costo medio di classe.

Riportiamo a titolo esemplificativo l'analisi effettuata per l'età, rimandando alla fine di questo capitolo per le tabelle e i grafici relativi alle altre variabili tariffarie.

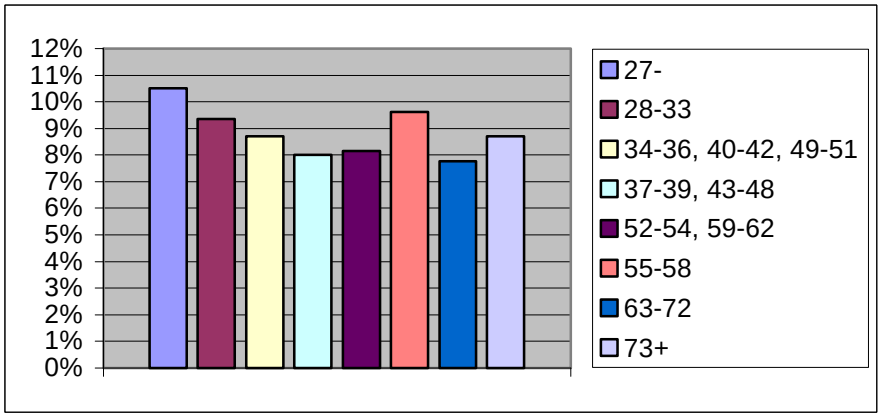
Età	Sinistri	Costo	Esposizione	Frequenza	Costo medio
27-	616	1941681,27	5860,51	0,10511029	3152,08
28-33	1131	2641185,83	12099,67	0,09347363	2335,27
34-36, 40-42, 49-51	1668	4540637,26	19134,76	0,08717121	2722,2
37-39, 43-48	1474	4069200,09	18400,01	0,08010866	2760,65
52-54, 59-62	1009	1996159,96	12366,51	0,08159136	1978,35
55-58	629	1510442,22	6542,38	0,09614237	2401,34
63-72	897	2000766,71	11531,46	0,07778718	2230,51
73+	473	846048,78	5440,05	0,08694769	1788,69

Tab. 4.1.1 Frequenza sinistri e costo medio per l'età

Nella tabella 4.1.1 sono riportati per ogni classe di età il numero totale di sinistri, il costo totale osservato nella classe, l'esposizione totale, la frequenza sinistri e infine il costo medio per sinistro espresso in Euro. Si osservi che la frequenza sinistri è abbastanza differenziata con l'età; questa osservazione ci fa capire che quando vaglieremo le variabili tariffarie significative per la frequenza sinistri l'età verrà selezionata. Si osservi che la frequenza relativa agli individui di età inferiore ai ventisette anni è molto elevata, mentre quella degli assicurati di età compresa tra i sessantatre e settantadue anni è decisamente minore.

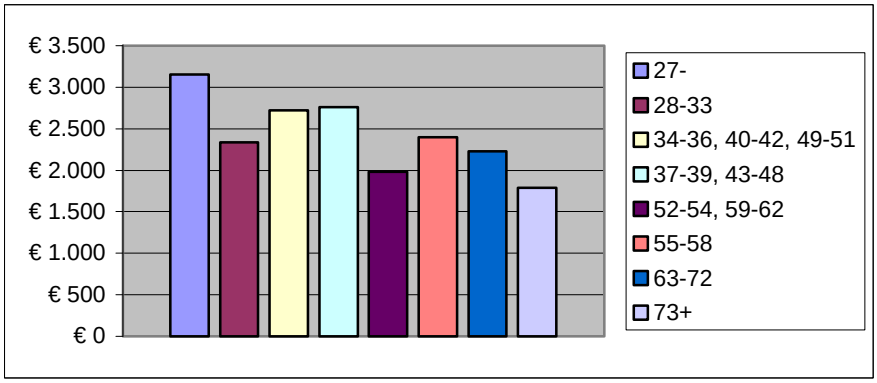
Anche il costo medio appare abbastanza differenziato, ed analogamente a quanto detto per la frequenza sinistri, c'è da aspettarsi che l'età risulti significativa per i costi.

Riassumiamo i dati presentati nella tabella 4.1.1 nei due grafici successivi.



Graf. 4.1.1 Frequenza sinistri in funzione dell'età

[4.1]



Graf. 4.1.2 Costo medio in funzione dell'età

4.2 I MODELLI LINEARI GENERALIZZATI

Per la stima dei modelli a priori, è stata utilizzata la procedura GENMOD del SAS che si avvale dei modelli lineari generalizzati. Diamo qui di seguito un breve cenno per richiamare la teoria che sta alla base delle valutazioni fatte.

Nei modelli lineari generalizzati si assume che la generica variabile risposta Y abbia una distribuzione di probabilità appartenente alla famiglia esponenziale. In altre parole la densità di probabilità per le variabili aleatorie continue, o la distribuzione di probabilità per le variabili discrete, può venire espressa nella forma

$$f(y) = \left(\frac{\exp yq - b(q)}{a(f)} + c(y, f) \right)$$

dove *a*, *b* e *c* sono funzioni che determinano la particolare distribuzione e c ha il ruolo di costante di normalizzazione. Nel seguito si assume $a(f) = \frac{f}{w}$ e $c = c(y, \frac{f}{w})$, dove w rappresenta un peso noto per ciascuna osservazione. La forma della distribuzione è allora individuata dalla funzione *b.*

Per questo tipo di distribuzioni la teoria fornisce le seguenti espressioni della speranza matematica e della varianza,

$$E(Y) = b'(q),$$

$$Var(Y) = b''(q)\frac{f}{w}.$$ [8]

Posto $m = E(Y)$, la varianza può essere espressa in funzione della media con:

$$Var(Y) = V(m)\frac{f}{w}$$

dove *V* è la *funzione di varianza*, $V(m) = b''\left(b'^{-1}(m)\right)$.

Le distribuzioni di probabilità della variabile risposta *Y* sono solitamente espresse in funzione della media m, e del *parametro di dispersione* f, e non in dipendenza dal *parametro canonico* q. Per alcune distribuzioni f è noto, mentre per altre deve essere stimato.

La media m_i della variabile risposta per l'*i*-esima osservazione è legata ad una componente lineare, $h_i = x_i'b$, attraverso una funzione differenziabile e monotona in senso stretto *g*, detta *funzione di collegamento*: $g(m_i) = h_i$. Il vettore x_i' rappresenta le caratteristiche dell'individuo *i* osservabili a priori, che sono le variabili esplicative, e b è un vettore di parametri da stimare, detti *parametri di regressione*.

Se $\hat{b}$ è la stima di b allora $\hat{m}_i = g^{-1}\left(x_i'\hat{b}\right)$ è una stima della speranza matematica dell'*i*-esima variabile risposta.

Per la stima di b si ricorre al metodo della massima verosimiglianza.

La procedura GENMOD del SAS utilizza l'algoritmo di Newton-Raphson per massimizzare la funzione di log-verosimiglianza $l(y,m,f)$ [9] rispetto ai

---

[8] Si osservi che in questo contesto, gli apici dopo un simbolo di funzione, indicano le derivate prima e seconda rispetto a q.

[9] I simboli y e m che compaiono nell'espressione di log-verosimiglianza rappresentano dei vettori.

parametri di regressione. Alla k-esima iterazione l'algoritmo aggiorna il vettore b_k con

$$b_{k+1} = b_k - \mathrm{H}^{-1}\mathrm{s}$$

dove H è la matrice Hessiana (o delle derivate seconde), e s rappresenta il vettore dei *gradienti* (derivate prime) della funzione di log-verosimiglianza entrambi valutati con il valore corrente del vettore di parametri.

La procedura GENMOD utilizza principalmente due statistiche per verificare la bontà di accostamento ai dati del modello stimato: *la devianza scalata* e la *statistica chi-quadrato del Pearson*.

Un modello è individuato dalla presenza di un certo insieme di variabili esplicative. Il vettore x_i individua le determinazioni o i livelli di queste variabili per l'*i*-esima osservazione. Il numero delle componenti di b è il numero di parametri da stimare. Il massimo numero di parametri da considerare è pari al numero delle osservazioni. Con questo modello detto *saturo* si ottengono stime $\hat{m}_i$ coincidenti con i valori osservati y_i.

Per un fissato valore del parametro di dispersione f, la devianza scalata è definita come due volte la differenza tra la massima log-verosimiglianza raggiungibile, che è quella del modello saturo, e la log-verosimiglianza che si ottiene utilizzando le stime di massima verosimiglianza dei parametri del modello specificato.

Se $l(y_i, \hat{m}_i)$ rappresenta la log-verosimiglianza per l'osservazione i espressa in funzione del valore stimato della speranza matematica $\hat{m}_i$ e della determinazione delle variabili risposta y_i allora la devianza scalata è

$$D^{\times}(y, m) = 2\sum_i \left(l(y_i, y_i) - l(y_i, \hat{m}_i)\right) \qquad (4.2.1)$$

La (4.2.1) si può esprimere come

$$D^{\times}(y,m) = \frac{D(y,m)}{f}$$

dove *D* rappresenta la devianza.

La statistica chi-quadrato del Pearson è definita come

$$c^2 = \sum_i \frac{w_i (y_i - m_i)^2}{V(m_i)},$$

e la statistica chi-quadrato del Pearson scalata è $\frac{c^2}{f}$.

La versione scalata di entrambe le statistiche, sotto certe condizioni di regolarità ha, asintoticamente, una distribuzione chi-quadrato con un numero di gradi di libertà pari al numero di osservazioni meno il numero di parametri stimati.

Consideriamo ora il caso di f non noto. In questi casi si può utilizzare una sua stima. Un suggerimento ci viene dal metodo dei momenti. Dato che la distribuzione asintotica della devianza scalata, $D^* = \frac{D}{f}$, è chi-quadrato con $n - p$ gradi di libertà, dove *n* è il numero di osservazioni e *p* è il numero di parametri, si ha $E(D^*) = n - p$. Uguagliando $n - p$ con la media empirica e risolvendo per f si ottiene $\hat{f} = \frac{D}{n-p}$, con *D* determinazione della devianza. Un altro suggerimento per la stima di f ci viene dalla statistica c^2 scalata per la quale la speranza matematica è pari a $n - p$. Dividendo la determinazione di c^2 per $n - p$ si ottiene la stima di f.

Per i modelli per cui f è stimato, per verificare la bontà del modello si utilizza la statistica F. Se D_0 è la devianza risultante da un modello e D_1 è la devianza di un sotto-modello, allora, sotto appropriate condizioni di regolarità e

indipendenza, la statistica $\frac{D_1 - D_0}{r\frac{D_0}{n-p}}$ ha distribuzione asintotica F di parametri r e $n-p$, dove r è la differenza tra i numeri di parametri dei due modelli.

L'output della procedura GENMOD del SAS si compone di varie parti. Ne riportiamo qui di seguito, a titolo illustrativo, il risultato ottenuto per un modello per i costi dei sinistri.

```
The SAS System                    09:07 Monday, October 8, 2001   1

                        The GENMOD Procedure

                          Model Information
Description                          Value

Data Set                             WORK.RCAX
Distribution                         GAMMA
Link Function                        LOG
Dependent Variable                   COSTO
Offset Variable                      LN
Observations Used                    7442
```

Tab. 4.2.1 *Model Information*

Nella tabella 4.2.1 vengono riportati il nome del data set utilizzato, la distribuzione scelta per la variabile risposta, la funzione di collegamento, il nome della variabile risposta utilizzata nel data set, la variabile *offset*, in questo caso il logaritmo del numero di sinistri, ed infine il numero di osservazioni utilizzate.

La seconda parte dell'output specifica le variabili utilizzate per le componenti di regressione.

```
                    Class Level Information

Class     Levels  Values

POT_EFF        6  32- 33-39 40-54 55-64, 67-76,
                  65-66 77-88 89+
PROV           6  prov1 prov2 prov3 prov4 prov5
                  prov6
ETA            8  27- 28-33 34-36, 40-42, 49
                  37-39, 43-48 52-54, 59-62
                  55-58 63-72 73+
```

Tab. 4.2.2 *Class Level Information*

Nella Class level information della tabella 4.2.2, vengono riportate le variabili del modello con a fianco il numero di classi o livelli e la loro descrizione.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7389	12196.8056	1.6507
Scaled Deviance	7389	1956.6420	0.2648
Pearson Chi-Square	7389	46059.6253	6.2335
Scaled Pearson X2	7389	7389.0000	1.0000
Log Likelihood	.	-69405.5734	.

Tab. 4.2.3 *Criteria For Assessing Goodness of Fit*

Per quanto riguarda invece i criteri utilizzati per valutare la bontà della stima, nella tabella 4.2.3, per la devianza, la devianza scalata, la statistica chi-quadrato, e la statistica chi-quadrato scalata vengono riportati i gradi di libertà, pari al numero di osservazioni meno il numero di variabili stimate, le determinazioni delle statistiche ed infine le determinazioni divise per i gradi di libertà. Questo output è ottenuto con l'opzione PSCALE che stima f con il rapporto Pearson c^2/DF. Nell'ultima riga della tabella viene riportato il valore della log-verosimiglianza.

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	7.8226	0.4104	363.3740	0.0001
POT_EFF	32-		1	-0.4144	0.1470	7.9487	0.0048
POT_EFF	33-39		1	-0.3107	0.1335	5.4145	0.0200
POT_EFF	40-54		1	-0.0229	0.1235	0.0343	0.8531
POT_EFF	55-64, 67-76		1	-0.0098	0.1315	0.0055	0.9408
POT_EFF	65-66, 77-88		1	-0.0622	0.1358	0.2100	0.6467
POT_EFF	89+		0	0.0000	0.0000	.	.
PROV	prov1		1	-0.0059	0.4717	0.0002	0.9901
PROV	prov2		1	-0.4865	0.4847	1.0076	0.3155
PROV	prov3		1	0.0196	0.4982	0.0015	0.9686
PROV	prov4		1	-0.6726	0.6483	1.0763	0.2995
PROV	prov5		1	-0.2265	0.4554	0.2474	0.6189
PROV	prov6		0	0.0000	0.0000	.	.
ETA	27-		1	-0.0471	0.5303	0.0079	0.9293
ETA	28-33		1	-0.2236	0.4577	0.2387	0.6251
ETA	34-36, 40-42, 49		1	0.1722	0.4466	0.1486	0.6999
ETA	37-39, 43-48		1	-0.1256	0.4604	0.0745	0.7849
ETA	52-54, 59-62		1	-0.6233	0.4643	1.8020	0.1795
ETA	55-58		1	-0.2411	0.4928	0.2394	0.6246
ETA	63-72		1	-0.5650	0.4716	1.4357	0.2308
ETA	73+		0	0.0000	0.0000	.	.
PROV*ETA	prov1	27-	1	0.7263	0.6155	1.3926	0.2380
PROV*ETA	prov1	28-33	1	0.3240	0.5417	0.3578	0.5497
PROV*ETA	prov1	34-36, 40-42, 49	1	0.3537	0.5269	0.4507	0.5020
PROV*ETA	prov1	37-39, 43-48	1	0.1788	0.5391	0.1100	0.7402
PROV*ETA	prov1	52-54, 59-62	1	0.9001	0.5539	2.6406	0.1042
PROV*ETA	prov1	55-58	1	0.1488	0.5943	0.0627	0.8023
PROV*ETA	prov1	63-72	1	0.5546	0.5621	0.9733	0.3239
PROV*ETA	prov1	73+	0	0.0000	0.0000	.	.
PROV*ETA	prov2	27-	1	0.7940	0.6391	1.5438	0.2140
PROV*ETA	prov2	28-33	1	0.7535	0.5613	1.8025	0.1794

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
PROV*ETA	prov2	34-36, 40-42, 49	1	0.2161	0.5445	0.1575	0.6915
PROV*ETA	prov2	37-39, 43-48	1	1.4540	0.5605	6.7288	0.0095
PROV*ETA	prov2	52-54, 59-62	1	0.9377	0.5687	2.7185	0.0992
PROV*ETA	prov2	55-58	1	0.6241	0.6110	1.0431	0.3071
PROV*ETA	prov2	63-72	1	1.1886	0.5801	4.1984	0.0405
PROV*ETA	prov2	73+	0	0.0000	0.0000	.	.
PROV*ETA	prov3	27-	1	0.0246	0.6559	0.0014	0.9701
PROV*ETA	prov3	28-33	1	0.3109	0.5769	0.2903	0.5900
PROV*ETA	prov3	34-36, 40-42, 49	1	-0.1591	0.5575	0.0815	0.7753
PROV*ETA	prov3	37-39, 43-48	1	-0.0090	0.5701	0.0002	0.9874
PROV*ETA	prov3	52-54, 59-62	1	0.3198	0.5875	0.2963	0.5862
PROV*ETA	prov3	55-58	1	0.4219	0.6325	0.4450	0.5047
PROV*ETA	prov3	63-72	1	0.2643	0.5986	0.1949	0.6589
PROV*ETA	prov3	73+	0	0.0000	0.0000	.	.
PROV*ETA	prov4	27-	1	0.3079	0.8544	0.1299	0.7186
PROV*ETA	prov4	28-33	1	0.7420	0.7510	0.9762	0.3231
PROV*ETA	prov4	34-36, 40-42, 49	1	0.6199	0.7240	0.7330	0.3919
PROV*ETA	prov4	37-39, 43-48	1	1.3093	0.7379	3.1479	0.0760
PROV*ETA	prov4	52-54, 59-62	1	1.1255	0.7624	2.1794	0.1399
PROV*ETA	prov4	55-58	1	0.9144	0.8800	1.0797	0.2988
PROV*ETA	prov4	63-72	1	1.8555	0.7720	5.7776	0.0162
PROV*ETA	prov4	73+	0	0.0000	0.0000	.	.
PROV*ETA	prov5	27-	1	0.6675	0.6249	1.1410	0.2854
PROV*ETA	prov5	28-33	1	0.5178	0.5339	0.9406	0.3321
PROV*ETA	prov5	34-36, 40-42, 49	1	0.1127	0.5136	0.0481	0.8263
PROV*ETA	prov5	37-39, 43-48	1	0.1309	0.5289	0.0612	0.8046
PROV*ETA	prov5	52-54, 59-62	1	0.6951	0.5399	1.6571	0.1980
PROV*ETA	prov5	55-58	1	0.7828	0.5744	1.8572	0.1729
PROV*ETA	prov5	63-72	1	0.8116	0.5477	2.1961	0.1384
PROV*ETA	prov5	73+	0	0.0000	0.0000	.	.
PROV*ETA	prov6	27-	0	0.0000	0.0000	.	.
PROV*ETA	prov6	28-33	0	0.0000	0.0000	.	.
PROV*ETA	prov6	34-36, 40-42, 49	0	0.0000	0.0000	.	.
PROV*ETA	prov6	37-39, 43-48	0	0.0000	0.0000	.	.
PROV*ETA	prov6	52-54, 59-62	0	0.0000	0.0000	.	.
PROV*ETA	prov6	55-58	0	0.0000	0.0000	.	.
PROV*ETA	prov6	63-72	0	0.0000	0.0000	.	.
PROV*ETA	prov6	73+	0	0.0000	0.0000	.	.
SCALE			0	0.1604	0.0000	.	.

NOTE: The Gamma scale parameter was estimated by DOF/Pearson's Chi-Squared

Tab. 4.2.4 *Analysis of Parameter Estimates*

Nella sezione analysis of parameter estimates, per ciascun parametro del modello vengono presentati il nome della variabile e il livello, i gradi di libertà, il valore stimato, lo standard error, il valore chi-quadrato di Wald che è dato da $\left(\frac{Estimate}{StdErr}\right)^2$ e che asintoticamente è una $c^2(1)$, e il *p-value* basato sulla distribuzione chi-quadrato. Nelle ultime due righe viene riportato il valore del parametro di scala, che nel caso gamma è $\frac{1}{f}$, e il metodo con cui tale valore è stato calcolato.

Concludiamo questo paragrafo presentado l'output con le statistiche utilizzate per la scelta dei modelli. Nella tabella 4.2.5, per il fattore di scala, viene presentato il valore della distribuzione c^2 con un grado di libertà, e il rispettivo *p*-value. Inoltre, in corrispondenza di ciascuna variabile, vengono riportati il nome, la devianza del modello che comprende la variabile e tutte le precedenti, i

gradi di libertà per il numeratore e per il denominatore, la statistica F per testare la significatività dell'effetto, il *p*-value per la statistica F, la statistica chi-quadrato data dalla differenza tra la devianza scalata di due modelli consecutivi, e il *p*-value della distribuzione chi-quadro con i gradi di libertà del numeratore.

Lagrange Multiplier Statistics

Parameter	ChiSquare	Pr>Chi
Scale	3375.9770	0.0001

LR Statistics For Type 1 Analysis

Source	Deviance	NDF	DDF	F	Pr>F	ChiSquare	Pr>Chi
INTERCEPT	13106.4601	0	7389	.	.	.	.
POT_EFF	12873.3352	5	7389	7.4797	0.0001	37.3985	0.0001
PROV	12733.8138	5	7389	4.4765	0.0004	22.3824	0.0004
ETA	12612.4950	7	7389	2.7803	0.0069	19.4623	0.0069
PROV*ETA	12196.8056	35	7389	1.9053	0.0010	66.6859	0.0010

Tab. 4.2.5 *Lagrange Multiplier Statistic* e *Likelihood Ratio*

4.3 I MODELLI BASE

Il metodo che abbiamo utilizzato nella scelta del modello, cioè nella scelta delle variabili da inserire è un metodo di tipo *forward.* Si parte dal *modello nullo* con la sola intercetta e si inseriscono via via nuove variabili. La scelta se inserire o meno una nuova variabile è fatta andando a controllare il *p*-value in corrispondenza della statistica del rapporto di verosimiglianza, e scegliendo di inserire la variabile in presenza di un p-value inferiore alla soglia prefissata del 1%.

Una volta inserite tutte le variabili significative, si è passati a considerare le interazioni tra le variabili già inserite, cioè le coppie di variabili, e ad inserire le significative.

Al termine di questa fase di inserimento si è passati ad un'analisi di tipo *backward* in modo da verificare se l'inserimento di qualche componente, variabile o coppia di variabili, risultasse superflua.

[4.3]

Abbiamo considerato per la frequenza sinistri il modello con distribuzione per la variabile risposta di tipo Poisson, e per i costi un modello con distribuzione gamma ed uno con distribuzione log-normale.

- *Modello di Poisson*

Per il numero di sinistri, si è adottata una distribuzione di Poisson con un modello di tipo moltiplicativo, cioè con funzione di collegamento logaritmica. La stima è stata fatta su 134655 osservazioni ciascuna delle quali è stata pesata con la durata del periodo di esposizione pari o inferiore ad un anno. Sono stati selezionati sei fattori tariffari:

- Provincia
- Età
- Anno
- Cavalli fiscali
- Immatricolazione
- Professione

e due interazioni,

- Provincia*Anno
- Provincia*Professione.

Tutte le variabili ed interazioni selezionate risultano molto significative nella spiegazione del numero di sinistri, infatti il *p*-value più alto riscontrato è 0,32%, e quindi ben inferiore alla soglia fissata.

Nella prossima tabella, i coefficienti (St. Coeff.) stimati per i diversi livelli dei fattori tariffari e poi standardizzati in modo che abbiano media unitaria, sono messi a confronto con la *relative severity* (Rel. Sev.) di ciascun livello, ossia con il rapporto tra la frequenza sinistri del livello e la frequenza sinistri globale.

FATTORE TARIFFARIO: Provincia

Livello	Peso(%)	St. Coeff.	Rel. Sev.
prov1	28,9%	0,772	0,884
prov2	17,2%	1,063	1,065
prov3	18,7%	1,013	0,873
prov4	7,1%	0,737	0,836
prov5	18,8%	1,228	1,227
prov6	9,2%	1,313	1,162

FATTORE TARIFFARIO: Età

Livello	Peso(%)	St. Coeff.	Rel. Sev.
27-	6,4%	1,254	1,216
28-33	13,2%	1,090	1,082
34-36, 40-42, 49-51	20,9%	1,013	1,009
37-39, 43-48	20,1%	0,931	0,927
52-54, 59-62	13,5%	0,936	0,944
55-58	7,2%	1,103	1,112
63-72	12,6%	0,883	0,900
73+	6,0%	0,982	1,006

FATTORE TARIFFARIO: Anno

Livello	Peso(%)	St. Coeff.	Rel. Sev.
1998	29,5%	1,134	0,889
1999	35,3%	0,936	1,043
2000	35,2%	0,951	1,050

FATTORE TARIFFARIO: Cavalli

Livello	Peso(%)	St. Coeff.	Rel. Sev.
11-	5,9%	0,773	0,803
12	16,8%	0,978	0,987
13-14	26,3%	1,035	1,039
15-17	30,5%	0,983	0,981
18-19	11,6%	1,047	1,025
20+	8,9%	1,085	1,072

FATTORE TARIFFARIO: Immatricolazione

Livello	Peso(%)	St. Coeff.	Rel. Sev.
1	4,1%	0,823	0,777
2, 3	6,8%	0,942	1,133
4	19,3%	0,981	0,972
5	12,5%	1,070	1,016
6-8	9,3%	0,930	1,078
9-10	11,1%	0,984	0,980
11-12	26,1%	1,056	0,977
13+	10,9%	0,997	1,041

FATTORE TARIFFARIO: Professione

Livello	Peso(%)	St. Coeff.	Rel. Sev.
1	45,8%	0,986	0,948
2	6,1%	1,144	1,123
3	4,8%	1,242	1,129
4	15,5%	0,750	1,027
5	9,2%	0,928	0,973
6	1,5%	0,707	0,955
7	5,5%	0,991	0,984
8	11,5%	1,314	1,087

Tab. 4.3.1 Fattori tariffari per la frequenza sinistri (modello di Poisson)

Si considerino, per esempio, i contraenti più giovani, cioè quelli di età inferiore ai ventisette anni, e quelli di età 63-72. Come ci si attendeva, per i primi il coefficiente standardizzato è ben superiore alla media, mentre per i secondi è minore.

Si prendano ora quegli individui che hanno determinazione 4 in corrispondenza della professione. Per tali rischi la relative severity è pari a 1,027 mentre il coefficiente standardizzato ricavato dal modello di Poisson è 0,750. Il rapporto 1,027/0,750=1,369 significa che se facessimo valutazioni in maniera indipendente dagli altri fattori tariffari, il coefficiente per la frequenza sinistri per questi contraenti risulterebbe maggiore di quasi il 37% rispetto a quello stimato nel nostro modello.

- *Modello Gamma*

Per il costo dei sinistri si è adottata una distribuzione Gamma con un modello di tipo moltiplicativo. La stima è stata fatta utilizzando 7442 osservazioni, ovvero si sono considerati quei contraenti che hanno riportato almeno un sinistro, ed inoltre ciascuna osservazione è stata pesata con il numero di sinistri.

Sono stati selezionati tre fattori tariffari:

- Potenza in kilowatt
- Provincia
- Età

e un'interazione,

- Provincia*Età.

Tutte le variabili ed interazioni selezionate risultano molto significative nella spiegazione dei costi dei sinistri, infatti il *p*-value più alto riscontrato è 0,69%. Nella prossima tabella, i coefficienti (St. Coeff.) collegati ai diversi livelli dei fattori tariffari sono standardizzati in modo da avere media unitaria, mentre la relative severity (Rel. Sev.) di ciascun livello è calcolata attraverso il rapporto tra il costo medio del livello e il costo medio globale.

FATTORE TARIFFARIO: Potenza

Livello	Peso(%)	St. Coeff.	Rel. Sev.
32-	9,7%	0,731	0,686
33-39	16,6%	0,811	0,769
40-54	32,6%	1,081	1,085
55-64, 67-76	19,1%	1,095	1,127
65-66, 77-88	15,1%	1,039	1,006
89+	6,8%	1,106	1,238

FATTORE TARIFFARIO: Provincia

Livello	Peso(%)	St. Coeff.	Rel. Sev.
prov1	25,6%	1,163	1,164
prov2	18,3%	0,719	1,131
prov3	16,3%	1,193	0,874
prov4	6,0%	0,597	1,066
prov5	23,1%	0,932	0,900
prov6	10,7%	1,170	0,755

FATTORE TARIFFARIO: Età

Livello	Peso(%)	St. Coeff.	Rel. Sev.
27-	7,8%	1,109	1,274
28-33	14,3%	0,930	0,943
34-36, 40-42, 49-51	21,1%	1,381	1,100
37-39, 43-48	18,7%	1,026	1,115
52-54, 59-62	12,8%	0,623	0,799
55-58	8,0%	0,914	0,970
63-72	11,4%	0,660	0,901
73+	6,0%	1,163	0,723

Tab. 4.3.2 Fattori tariffari per il costo medio (modello Gamma)

Le considerazioni che si possono fare sono analoghe a quelli effettuate per il modello di Poisson.

- *Modello Log-normale*

Per il costo dei sinistri in alternativa alla distribuzione Gamma abbiamo considerato la distribuzione normale per il logaritmo dei costi con un modello di tipo addittivo, cioè con funzione di collegamento identica.

Per la stima di questo modello è necessario disporre dei risarcimenti per singolo sinistro, mentre i dati a nostra disposizione presentano solo il risarcimento globale per polizza. Abbiamo quindi deciso di suddividere, in qualche modo, il totale per poter esemplificare anche l'uso di questo modello.

Il metodo che abbiamo utilizzato per scorporare i costi si basa sulla considerazione che, per quei rischi che hanno riportato più di un sinistro, un risarcimento elevato è verosimilmente il risultato della somma di un sinistro con

costo elevato ed altri con costi modesti. Pertanto si è deciso di operare nel seguente modo:

n_i rappresenti il numero di sinistri per il rischio i-esimo, con $n_i \geq 2$, c_i sia il risarcimento globale per la polizza *i*, e $\overline{c}$ sia il costo medio osservato per l'intero portafoglio. Se $c_i > n_i \overline{c}$, allora si è posto $\hat{c}_1 = \hat{c}_2 = ... = \hat{c}_{n_i - 1} = \overline{c}$ e $\hat{c}_{n_i} = c_i - \sum_{h=1}^{n_i - 1} \hat{c}_h$ mentre se $c_i \leq n_i \overline{c}$ riesce $\hat{c}_h = \frac{c_i}{n_i}$, con $h = 1,..., n_i$.

Dopo aver assegnato ad ogni individuo una variabile identificativa ed avere scorporato i costi, si è proceduto nella stima del modello. Si sono utilizzate 7897 osservazioni, tante quante il numero di sinistri. I fattori tariffari selezionati sono:

- Potenza in kilowatt
- Età.

Poiché entrambi i *p*-value sono inferiori a 0,01% si conclude che le due variabili selezionate risultano molto significative nella descrizione del logaritmo dei costi

Nella tabella 4.3.3, come per i casi Poisson e Gamma, vengono presentati i coefficienti standardizzati (St. Coeff.) collegati ai diversi livelli dei fattori tariffari e la relative severity (Rel. Sev.) di ciascun livello. In questo caso però, i coefficienti sono standardizzati in modo da avere media nulla e le relative severity sono ottenute dalla differenza tra la media dei logaritmi dei costi del livello e la media globale.

FATTORE TARIFFARIO: Età

Livello	Peso(%)	St. Coeff.	Rel. Sev.
27-	7,8%	0,025	0,171
28-33	14,3%	0,066	0,071
34-36, 40-42, 49-51	21,1%	0,152	0,032
37-39, 43-48	18,7%	0,056	-0,033
52-54, 59-62	12,8%	-0,051	-0,051
55-58	8,0%	-0,127	-0,007
63-72	11,4%	-0,086	-0,079
73+	6,0%	-0,458	-0,136

FATTORE TARIFFARIO: Potenza

Livello	Peso(%)	St. Coeff.	Rel. Sev.
32-	9,7%	-0,046	-0,089
33-39	16,6%	-0,109	-0,082
40-54	32,6%	0,004	0,038
55-64, 67-76	19,1%	0,024	0,004
65-66, 77-88	15,1%	0,024	0,013
89+	6,8%	0,193	0,109

Tab. 4.3.2 Fattori tariffari per il costo medio (modello Log-normale)

Per dare un idea dell'accostamento ai dati dei due modelli presentati per i costi si è presa in considerazione la radice quadrata della somma del quadrato dei residui.

Per il modello gamma riesce $\sqrt{\sum_{i,t}(c_{it}-\hat{c}_{it})^2}=240627{,}16$ mentre per il modello log-normale si ha $\sqrt{\sum_{i,t}(\log c_{it}-\hat{m}_{it})^2}=82{,}71$.

Conclusa la stima dei modelli a priori, utilizzando le espressioni degli stimatori presentate nel Capitolo 3, siamo passati alla stima dei modelli con eterogeneità descritti al Capitolo 1.

Nei prossimi paragrafi presentiamo inoltre alcune tabelle con i coefficienti bonus-malus per il numero di sinistri, i costi, ed il premio puro.

4.4 IL MODELLO DI POISSON CON ETEROGENEITÀ

In questo paragrafo presentiamo alcuni risultati ottenuti per il modello di Poisson ed in particolare si mostrerà che per i dati a nostra disposizione l'eterogeneità non dipende dal tempo.

Il modello di partenza è il modello a priori presentato nel paragrafo precedente. Si hanno 44885 assicurati osservati per tre periodi di tempo successivi ed i fattori tariffari che si sono utilizzati sono:

- Caratteristiche del veicolo: cavalli fiscali, immatricolazione.
- Caratteristiche del contraente: età, professione e provincia di residenza.
- Altri fattori tariffari: anno di osservazione, e le interazioni provincia-anno, provincia-professione.

Dalla stima a priori del modello di Poisson si ha:

$$\sum_{i,t}\left(n_{it}-\hat{l}_{it}\right)^2 = 8290,55;$$

$$\sum_{i,t}\hat{l}_{it}^2 = 612,61;$$

$$\sum_{i} n_i = 7897;$$

$$\sum_{i}\left(n_i-\hat{l}_i\right)^2 = 8956,67;$$

$$\sum_{i}\hat{l}_i^2 = 1592,83.$$

Se la componente di eterogeneità dipende dal tempo allora devono essere verificate le due condizioni presentate nel Paragrafo (3.6). Ovvero

$\widehat{s_r^2}, \widehat{s_r^2}^1 > 0$ se e solo se $\sum_i \sum_{t\neq t'}\left(n_{it}-\hat{l}_{it}\right)\left(n_{it'}-\hat{l}_{it'}\right) > 0$ e

$\widehat{s}_s^{\,2}, \widehat{s}_s^{\,2\,1} > 0$ se e solo se $\dfrac{\sum_{i,t}\left[\left(n_{it} - \hat{l}_{it}\right)^2 - n_{it}\right]}{\sum_{i,t}\hat{l}_{it}^2} > \dfrac{\sum_{i}\left[\left(n_{i} - \hat{l}_{i}\right)^2 - n_{i}\right]}{\sum_{i}\hat{l}_{i}^2}$.

Per i dati a nostra disposizione riesce:

$$\sum_i \sum_{t \neq t'} \left(n_{it} - \hat{l}_{it}\right)\left(n_{it'} - \hat{l}_{it'}\right) = 8956,67 - 8290,55 = 666,12 > 0$$

$$\frac{\sum_{i,t}\left[\left(n_{it} - \hat{l}_{it}\right)^2 - n_{it}\right]}{\sum_{i,t}\hat{l}_{it}^2} = \frac{393,55}{612,61} = 0,642$$

$$\frac{\sum_{i}\left[\left(n_{i} - \hat{l}_{i}\right)^2 - n_{i}\right]}{\sum_{i}\hat{l}_{i}^2} = \frac{1059,67}{1592,83} = 0,665.$$

Dunque la seconda condizione non è verificata, e si deve perciò concludere che la componente di eterogeneità per i dati a nostra disposizione non dipende dal tempo.

Allora la stima di s^2, riesce:

$$\hat{s}^2 = \frac{\sum_i nres_i^2 - \sum_i n_i}{\sum_i \hat{l}_i^2} = \frac{\sum_i \left(n_i - \hat{l}_i\right)^2 - \sum_i n_i}{\sum_i \hat{l}_i^2} = \frac{1059,67}{1592,83} = 0,665.$$

Concludiamo questo paragrafo riportando le tabelle con i coefficienti bonus-malus per la frequenza sinistri ottenuti attraverso la revisione bayesiana e l'approccio di credibilità.

		numero osservato di sinistri					
		0	1	2	3	4	5
premio per la frequenza	0,05	0,968	1,611	2,255	2,899	3,542	4,186
	0,1	0,938	1,561	2,185	2,808	3,432	4,055
	0,2	0,883	1,470	2,056	2,643	3,230	3,817
	0,5	0,750	1,250	1,749	2,248	2,747	3,246
	1	0,601	1	1,399	1,799	2,198	2,598
	2	0,429	0,715	1	1,285	1,571	1,856

Tab. 4.4.1 Coefficienti bonus-malus per la frequenza sinistri (revisione bayesiana)

		numero osservato di sinistri					
		0	1	2	3	4	5
premio per la frequenza	0,05	0,968	1	1,032	1,064	1,097	1,129
	0,1	0,938	1	1,062	1,125	1,187	1,249
	0,2	0,883	1	1,117	1,235	1,352	1,470
	0,5	0,750	1	1,250	1,499	1,749	1,998
	1	0,601	1	1,399	1,799	2,198	2,598
	2	0,429	1	1,571	2,142	2,712	3,283

Tab. 4.4.2 Coefficienti bonus-malus per la frequenza sinistri (credibilità lineare)

Si osservi che i due metodi conducono allo stesso risultato nel caso in cui l'assicurato non abbia riportato sinistri, mentre negli altri casi conducono a risultati significativamente diversi. In particolare si osservi che nell'approccio di tipo bayesiano il coefficiente bonus-malus riesce crescente al crescere del numero di sinistri, e decrescente all'aumentare del premio per la frequenza. Il coefficiente ricavato nell'approccio di credibilità lineare, invece, è funzione crescente della frequenza sinistri stimata, ed inoltre, all'aumentare del numero di sinistri, cresce più lentamente rispetto al coefficiente ricavato nell'approccio puramente bayesiano.

4.5 IL MODELLO CON ETEROGENEITÀ PER I COSTI

In Pinquet (1997) viene riportata l'espressione degli stimatori consistenti di $\hat{s}_U^2$ e $\hat{s}^2$ nel caso in cui si sia adottata una distribuzione di tipo log-normale.

Le espressioni degli stimatori sono:

$$\hat{s}_U^2 = \frac{\sum_{i:n_i \geq 2} \sum_{\substack{j,k \leq n_i \\ j \neq k}} lcres_{ij} lcres_{ik}}{\sum_i n_i (n_i - 1)}, \text{ e}$$

$$\hat{s}^2 = \widehat{s^2}^0 - \hat{s}_U^2 .$$

Per i dati a nostra disposizione riesce:

$$\sum_{i:n_i \geq 2} \sum_{\substack{j,k \leq n_i \\ j \neq k}} lcres_{ij} lcres_{ik} = 693,51;$$

$$\sum_i n_i (n_i - 1) = 2588;$$

$$\widehat{s^2}^0 = 1,313.$$

Dunque le stime di $\hat{s}_U^2$ e $\hat{s}^2$ sono

$$\hat{s}_U^2 = \frac{693,51}{2588} = 0,268$$

$$\hat{s}^2 = 1,313 - 0,268 = 1,045 .$$

Nella prossima tabella riportiamo i valori del coefficiente bonus-malus in funzione del numero di sinistri e del residuo *lcres* descritto nel Paragrafo 2.3, *Esempio 4*. I tre valori riportati nella colonna *lcres* corrispondono rispettivamente ai valori 0,5, 1 e 2 del rapporto costo/costo stimato.

[4.5]

	numero osservato di sinistri				
lcres	1	2	3	4	5
-0,69315	0,845	0,850	0,853	0,856	0,858
0	0,973	0,956	0,943	0,934	0,927
0,69315	1,121	1,075	1,043	1,020	1,003

Tab. 4.5.1 Coefficienti bonus-malus per la frequenza sinistri (credibilità lineare)

Consideriamo il caso in cui l'assicurato abbia riportato un sinistro. Se il valore del rapporto costo/costo stimato è 0,5 allora si ha un bonus per il costo del 15,5%, mentre se il rapporto tra i costi è 2, siamo in presenza di un malus per il costo del 12,1%.

4.6 IL MODELLO CON ETEROGENEITÀ PER IL PREMIO PURO

Come si è già detto i dati a nostra disposizione riguardano 44885 assicurati, i quali hanno riportato 7897 sinistri. La durata media dei periodi di osservazione è di circa otto mesi, e la frequenza sinistri a livello di portafoglio è del 8,6%. I fattori tariffari che si sono utilizzati sono, per la componente sinistri quelli selezionati per il modello di Poisson, e per la componente costi quelli utilizzati nel modello log-normale.

In questo paragrafo intendiamo stimare una distribuzione congiunta per i random effects relativi al numero e al costo dei sinistri attraverso la stima del modello con eterogeneità descritto nel Paragrafo 1.6.

Dall'output della procedura GENMOD del SAS, per il modello log-normale per il costo dei sinistri, si ricava

$$\widehat{s^2}^0 = 1,313.$$

Le statistiche necessarie per ottenere le stime consistenti sono:

$$\sum_i n_i = 7897;$$

$$\sum_i n_i(n_i - 1) = 2588$$

$$\sum_i \left(n_i - \hat{l}_i\right)^2 - n_i = 1059,67;$$

$$\sum_i \hat{l}_i^2 = 1592,83;$$

$$\sum_i \left(n_i - \hat{l}_i\right)\left(tlc_i - \widehat{tlc}_i\right) = 99,75;$$

$$\sum_i \left[\left(tlc_i - \widehat{tlc}_i\right)^2 - n_i \widehat{s^2}^0\right] = \sum_{i:n_i \geq 2} \sum_{\substack{j,k \leq n_i \\ j \neq k}} lcres_{ij} lcres_{ik} = 693,51.$$

Possiamo dunque ricavare le stime di $\hat{V}_{nn}^1$, $\hat{V}_{nn}$, $\hat{V}_{cn}$ e $\hat{V}_{cc}$.

$$\hat{V}_{nn}^1 = \frac{\sum_i \left(n_i - \hat{l}_i\right)^2 - n_i}{\sum_i \hat{l}_i^2} = 0,665;$$

$$\hat{V}_{cn} = \frac{\sum_i \left(n_i - \hat{l}_i\right)\left(tlc_i - \widehat{tlc}_i\right)}{\left(\sum_i \hat{l}_i^2\right)\left(1 + \hat{V}_{nn}^1\right)} = 0,038;$$

$$\hat{V}_{cc} = \frac{\sum_i \left[\left(tlc_i - \widehat{tlc}_i\right)^2 - n_i \widehat{s^2}^0\right]}{\left(\sum_i \hat{l}_i^2\right)\left(1 + \hat{V}_{nn}^1\right)} - \hat{V}_{cn}^2 = 0,260;$$

$$\hat{V}_{nn} = \log\left(1 + \hat{V}_{nn}^1\right) = 0,510;$$

Dalle stime ottenute si può calcolare l'indice di correlazione per i due effetti, riesce

$$\hat{r}_{cn} = \frac{\hat{V}_{cn}}{\sqrt{\hat{V}_{nn}\hat{V}_{cc}}} = 0,104,$$

da cui si deduce che la correlazione tra i due random effects è positiva.

Ricordiamo che, nell'approccio della credibilità lineare, il coefficiente bonus-malus per il premio puro per l'assicurato *i*, è dato da

[4.6]

$$1+a_{ni}(n_i-\hat{l}_i)+a_{ci}(tlc_i-\widehat{tlc_i})\,.$$

Con i valori numerici calcolati in precedenza, si ricavano:

$$m^i_{nn}=1+\hat{l}_i\hat{V}^1_{nn}=1+0,665\,\hat{l}_i\,,$$

$$m^i_{nc}=\hat{l}_i\hat{V}_{cn}(1+\hat{V}^1_{nn})=0,063\,\hat{l}_i\,,$$

$$b_n=\exp(\hat{V}_{nn}+\hat{V}_{cn})-1=0,730\,,$$

$$m^i_{cn}=m^i_{nc}=0,063\,\hat{l}_i\,,$$

$$m^i_{cc}=\hat{s}^{2^0}+\left[\hat{l}_i(1+\hat{V}^1_{nn})(\hat{V}^2_{cn}+\hat{V}_{cc})\right]=1,313+0,435\,\hat{l}_i\,,$$

$$b_c=(\hat{V}_{cn}+\hat{V}_{cc})\exp(\hat{V}_{nn}+\hat{V}_{cn})=0,515\,,$$

ed infine le soluzioni del sistema lineare presentato nel Paragrafo 2.5:

$$a_{ni}=\frac{m^i_{cc}b_n-m^i_{nc}b_c}{m^i_{nn}m^i_{cc}-\left(m^i_{nc}\right)^2}=\frac{0,958+0,285\,\hat{l}_i}{1,313+1,308\,\hat{l}_i-0,285\,\hat{l}_i^2}\,;$$

$$a_c=\frac{m^i_{nn}b_c-m^i_{nc}b_n}{m^i_{nn}m^i_{cc}-\left(m^i_{nc}\right)^2}=\frac{0,515+0,296\,\hat{l}_i}{1,313+1,308\,\hat{l}_i-0,285\,\hat{l}_i^2}\,.$$

Se l'assicurato non ha riportato alcun sinistro, allora il residuo per i costi è nullo, e il bonus è pari a

$$1-a_{ni}\,\hat{l}_i=\frac{1,313+0,350\,\hat{l}_i}{1,313+1,308\,\hat{l}_i-0,285\,\hat{l}_i^2}\,.$$

Nella prossima tabella presentiamo il coefficiente bonus malus calcolato attraverso l'approccio della credibilità lineare, al variare del residuo dei costi e del numero di sinistri riportati. con un premio per la frequenza unitario.

[4.6]

	numero di sinistri					
lcres	0	1	2	3	4	5
-1		0,721	1,148	1,575	2,002	2,429
-0,5		0,861	1,288	1,715	2,142	2,569
0	0,573	1	1,427	1,854	2,281	2,708
0,5		1,140	1,567	1,994	2,421	2,848
1		1,279	1,706	2,133	2,560	2,987

Tab. 4.6.1 Coefficienti bonus-malus per il premio puro (approccio di credibilità lineare)

Consideriamo il caso in cui l'assicurato abbia riportato un sinistro. Se il valore del residuo *lcres* è -0,5, allora si ha un bonus per il premio puro del 13,9%, viceversa se *lcres* è pari a 0,5, allora si è in presenza di un malus del 14%.

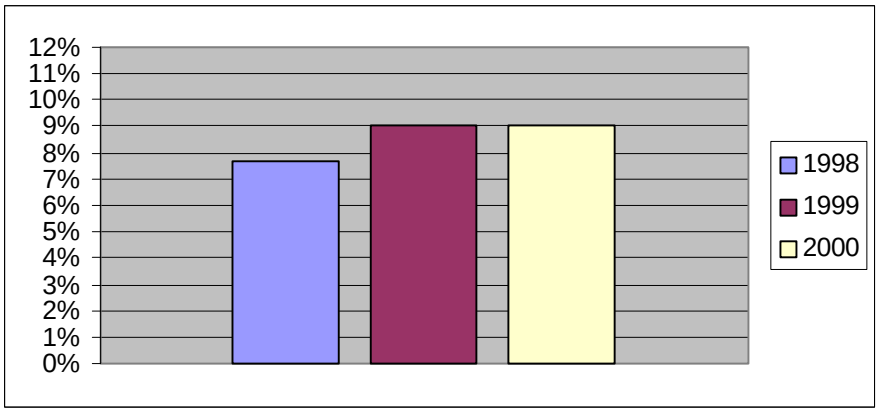
4.7 GRAFICI E TABELLE

Concludiamo questo capitolo con la presentazione dei grafici e delle tabelle.

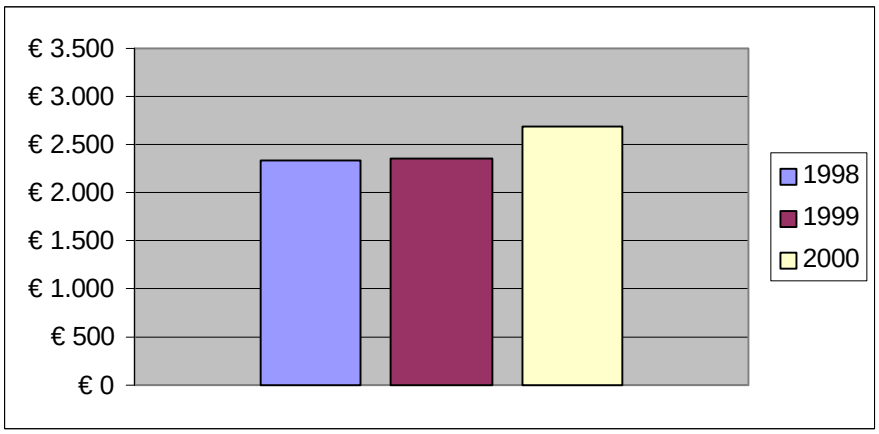
- *Variabile tariffaria:* ***Anno***

Anno	Sinistri	Costo	Esposizione	Frequenza	Costo medio
1998	2072	4836677,98	26956,03	0,07686593	2334,30
1999	2904	6837508,40	32224,67	0,09011729	2354,51
2000	2921	7871935,73	32194,65	0,09072935	2694,95

Tab. 4.7.1 Frequenza sinistri e costo medio per l'anno



Graf. 4.7.1 Frequenza sinistri in funzione dell'anno di osservazione

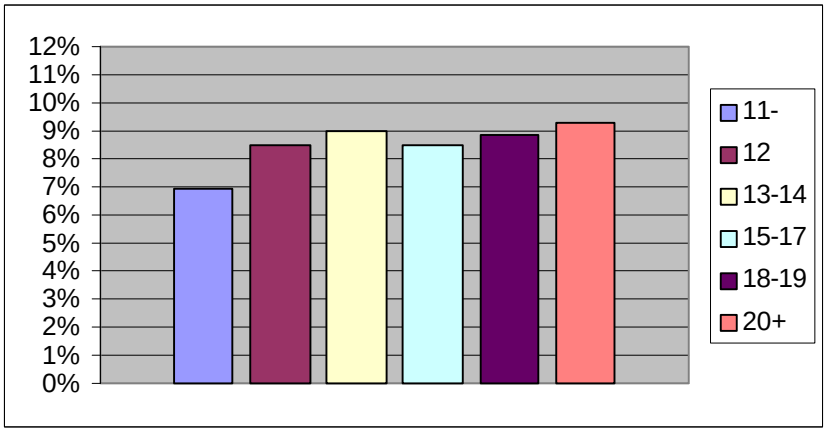


Graf. 4.7.2 Costo medio in funzione dell'anno di osservazione

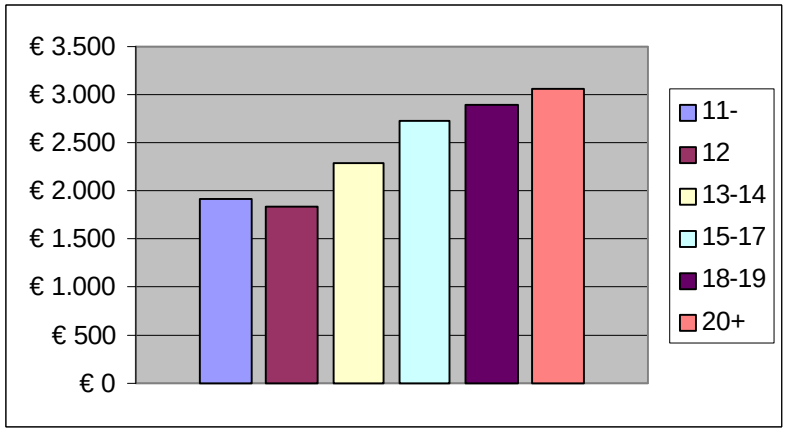
- *Variabile tariffaria:* ***Cavalli fiscali***

Cavalli	Sinistri	Costo	Esposizione	Frequenza	Costo medio
11-	374	716261,12	5386,31	0,06943524	1915,14
12	1305	2392822,09	15306,34	0,08525881	1833,58
13-14	2161	4941285,31	24056,63	0,08982971	2286,57
15-17	2364	6464038,54	27893,76	0,08475013	2734,36
18-19	942	2727176,57	10629,09	0,08862474	2895,09
20+	751	2304538,48	8103,22	0,09267915	3068,63

Tab. 4.7.2 Frequenza sinistri e costo medio per i cavalli fiscali



Graf. 4.7.3 Frequenza sinistri in funzione dei cavalli fiscali

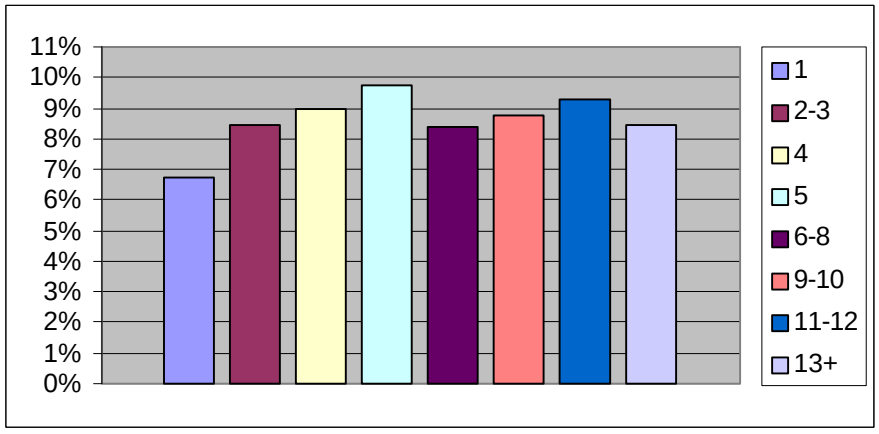


Graf. 4.7.4 Costo medio in funzione dei cavalli fiscali

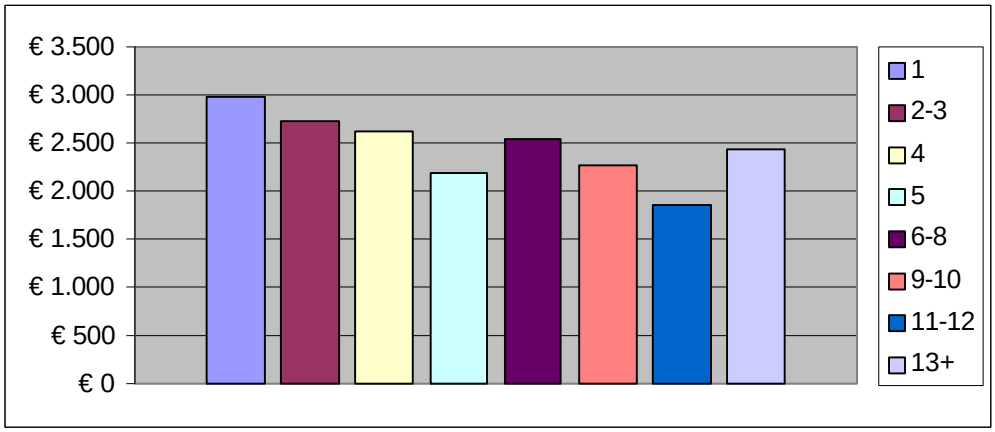
- *Variabile tariffaria:* ***Immatricolazione***

Immatricolazione	Sinistri	Costo	Esposizione	Frequenza	Costo medio
1	250	747081,34	3724,21	0,06712826	2988,33
2-3	2014	5513749,87	23862,71	0,08439945	2737,71
4	895	2352392,58	9944,22	0,09000201	2628,37
5	608	1333445,74	6211,46	0,09788357	2193,17
6-8	1481	3778207,91	17621,77	0,08404379	2551,12
9-10	1002	2273926,06	11415,07	0,08777868	2269,39
11-12	789	1459690,79	8465,83	0,09319821	1850,05
13+	858	2087627,81	10130,07	0,08469831	2433,13

Tab. 4.7.3 Frequenza sinistri e costo medio per l'anno di immatricolazione



Graf. 4.7.5 Frequenza sinistri in funzione dell'anno di immatricolazione

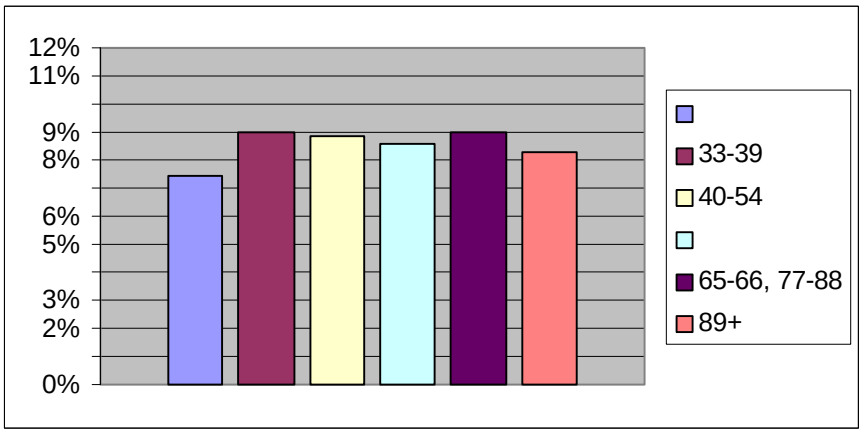


Graf. 4.7.6 Costo medio in funzione dell'anno di immatricolazione

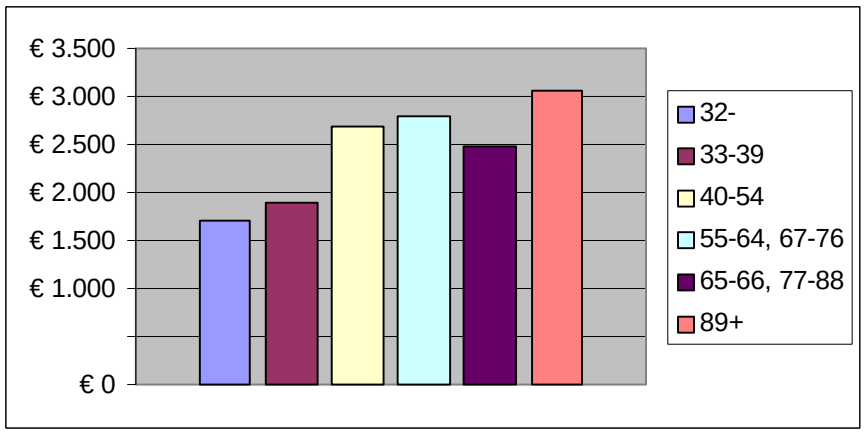
- *Variabile tariffaria:* ***Potenza in kilowatt***

Potenza	Sinistri	Costo	Esposizione	Frequenza	Costo Medio
32-	769	1305846,12	10363,88	0,07420005	1698,11
33-39	1313	2499812,42	14545,91	0,09026593	1903,89
40-54	2572	6907481,20	29115,32	0,08833837	2685,65
55-64, 67-76	1511	4213540,87	17575,50	0,08597196	2788,58
65-66, 77-88	1196	2977000,44	13321,51	0,08977959	2489,13
89+	536	1642441,06	6453,24	0,0830591	3064,26

Tab. 4.7.4 Frequenza sinistri e costo medio per la potenza in kilowatt



Graf. 4.7.7 Frequenza sinistri in funzione della potenza in kilowatt



Graf. 4.7.8 Costo medio in funzione della potenza in kilowatt

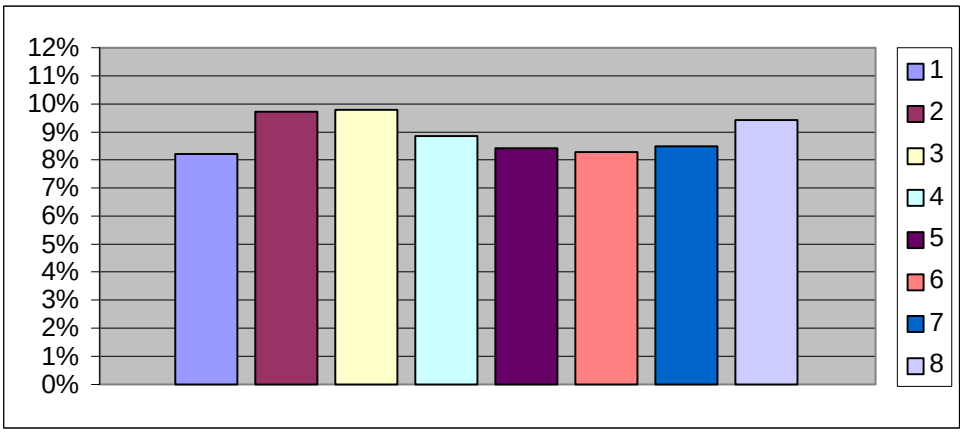
- *Variabile tariffaria:* ***Professione***

Codice	Professione
	Lavoratore dipendente
2	
3	Professioni che richiedono l'uso del veicolo
	Lavoratore del settore privato
5	
6	Lavoratore del settore terziario
	Pensio
8	Studente, casalinga, disoccupato

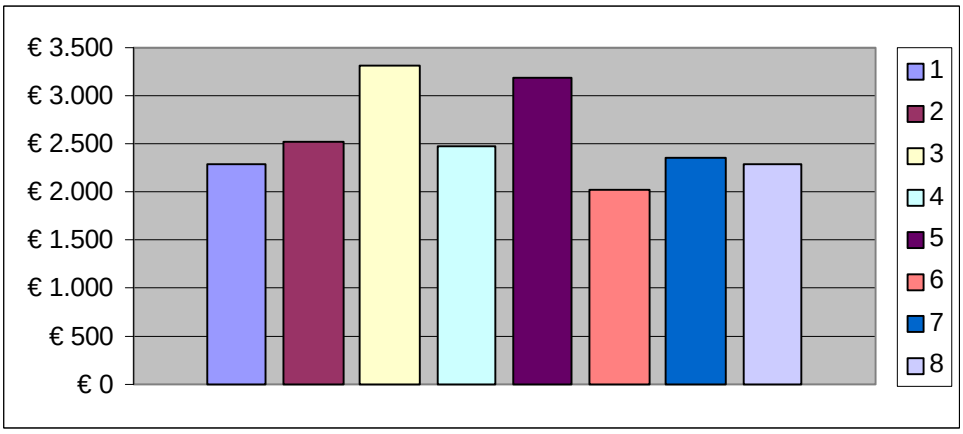
Codici professioni

	Sinistri	Costo		Frequenza	Costo medio
	3428	7878456,44		0,08192515	2298,27
	544	1373153,47		0,09701754	2524,18
	427	1412502,15		0,0976057	3307,97
	1260	3114208,40		0,0887363	2471,59
	710	2263779,25		0,08406742	3188,42
	112	225789,58		0,08256357	2015,98
	428	1006786,55		0,08500492	2352,31
	988	2271446,27		0,09397183	2299,03

Tab. 4.7.6



Graf. 4.7.9 Frequenza sinistri in funzione della professione



Graf. 4.7.10 Costo medio in funzione della professione

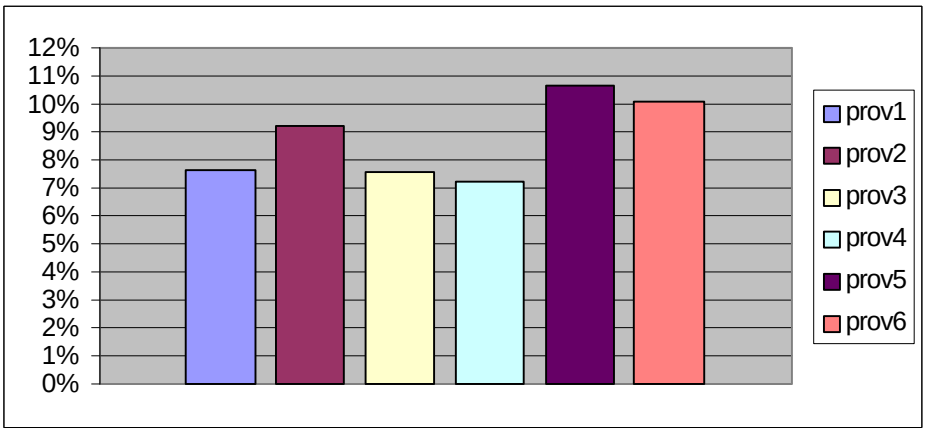
- *Variabile tariffaria:* ***Provincia di residenza***

Macro-provincia	**Province** *(sigla automobilistica)*
Prov.1	BG, BN, BR, BS, CB, CZ, EE, EN, FE, GE, GO, GR, IM, MN, MT, NO, PI, PS, PV, PZ, RE, SA, SI, SO, SR, SV, TN, TR, TV, UD, VA, VB, VC, VE
Prov.2	AP, AV, CA, FI, LC, ME, MI, MO, PA, PG, SS, VR
Prov.3	AL, AN, BI, BL, BZ, CH, CN, FO, FR, IS, LE, LI, LO, NU, OR, PD, PN, RC, RI, RO, TO, VI, VV
Prov.4	AG, AQ, AR, CE, CR, CS, FG, KR, PC, PE, RA, RG, RN, RS, SP, VT
Prov.5	AT, BO, LU, MS, NA, PO, RM, TA, TP, TS
Prov.6	AO, BA, CL, CO, LT, MC, PR, PT, TE

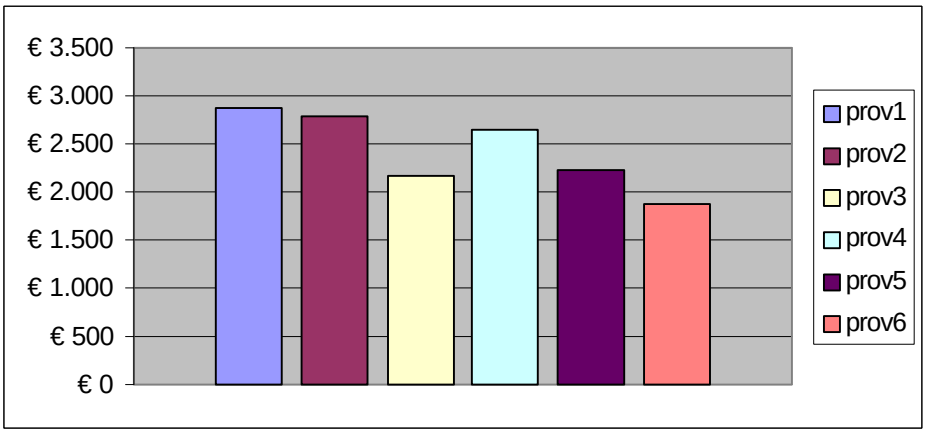
Tab.4.7.7 Classificazione della provincia di residenza

Provincia	Sinistri	Costo	Esposizione	Frequenza	Costo medio
prov1	2021	5822990,01	26442,83	0,07642903	2881,24
prov2	1447	4050372,20	15725,48	0,09201628	2799,15
prov3	1289	2788446,50	17081,37	0,07546232	2163,26
prov4	471	1243215,63	6520,14	0,07223772	2639,52
prov5	1827	4068006,65	17223,09	0,1060785	2226,6
prov6	842	1573091,11	8382,44	0,10044813	1868,28

Tab. 4.7.8 Frequenza sinistri e costo medio per la provincia di residenza



Graf. 4.7.11 Frequenza sinistri in funzione della provincia di residenza

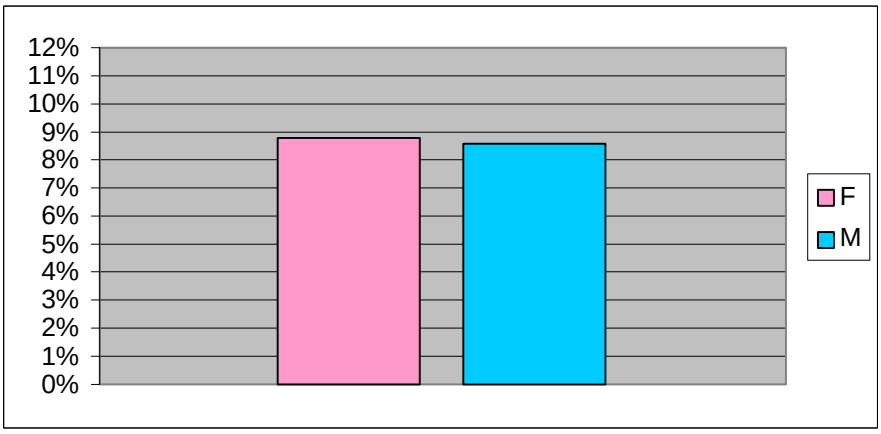


Graf. 4.7.12 Costo medio in funzione della provincia di residenza

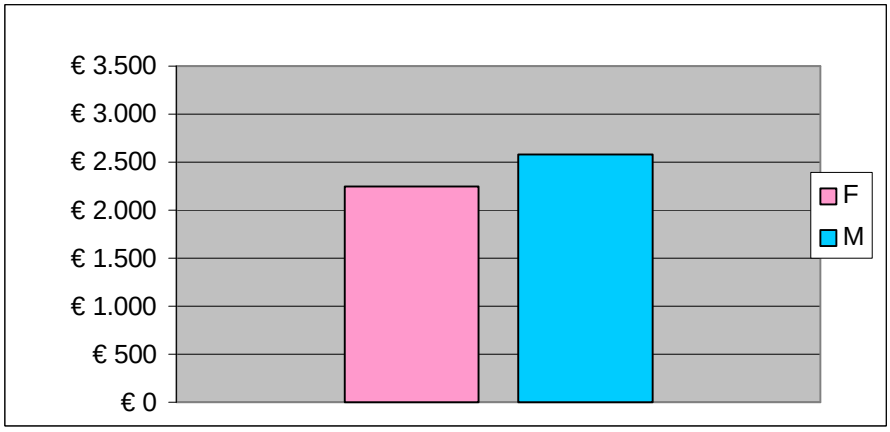
- *Variabile tariffaria:* ***Sesso***

Sesso	Sinistri	Costo	Esposizione	Frequenza	Costo medio
F	2668	6017361,97	30395,08	0,08777738	2255,38
M	5229	13528760,14	60980,27	0,08574904	2587,26

Tab. 4.7.9 Frequenza sinistri e costo medio per il sesso



Graf. 4.7.13 Frequenza sinistri in funzione del sesso



Graf. 4.7.14 Costo medio in funzione del sesso

BIBLIOGRAFIA

[1] Albrecht, P. (1982). "Parametric multiple regression risk models: Theory and statistical analysis". *Insurance: Mathematics and Economics 2, 49-46.* North-Holland Publishing Company.

[2] Cameron, A.C., and P.K. Trivedi (1998). *"Regression Analysis of Count Data"*. Econometric Society Monographs, Cambridge University Press.

[3] Hausman J.A., B.H. Hall, and Z. Griliches (1984). "Econometric Models for Count Data with an Application to the Patents-R&D Relationship". *Econometrica 52, 909-938.*

[4] Klugman S.A., F.H: Panjer, and G:E: Wilmot (1988). *"Loss Models. From Data to Decisions"*. Wiley Series in Probability and Statistics.

[5] Lemaire, J. (1977). "La Soif du Bonus". *ASTIN Bullettin 9, 181-190.*

[6] Lemaire, J. (1985). *"Automobile Insurance: Actuarial Models"*. Huebner International Series on Risk, Insurance and Econometric Security.

[7] McCullagh P. and J.A. Nelder (1989). *"Generalized Linear Models"*. London: Chapman and Hall.

[8] Pinquet J. (1996). "Hétérogénéité Inexpliquée". *Document de Travail THEMA 9714.*

[9] Pinquet, J. (1997). "Allowance for Cost of Claims in Bonus-Malus Systems". *ASTIN Bullettin 28, No. 2, 205-220.*

[10] Pinquet, J. (2000). "Experience rating for fleet of vehicles". *ASTIN Congress at Porto Cervo,* and forthcoming in ASTIN Bulletin.

[11] Pinquet, J. (2001a). “Experience rating through heterogeneous models”. THEMA, University Paris X, 92001 Nanterre, France.
Forthcoming in *Handbook of Insurance* (Chapter 15), Kluwer Academic Publishers. Huebner International Series on Risk, Insurance and Econometric Security.

[12] Pinquet, J. (2001b). “Linear credibility predictors for the pure premium of an insurance contract”. THEMA, University Paris X, 92001 Nanterre, France.

[13] Port, S. C. (1994). *“Theoretical probability for applications”*. Wiley Series in probability and mathematical statistics.

[14] Savron, I. (2001). “*Modelli con componenti di regressione per i numeri e per i costi dei sinistri*”. Tesi di laurea.