

5. I modelli lineari generalizzati

modello PEAR. LINEARE NORMALE → $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = E(Y_i) = \beta_0 + \sum_{k=1}^m \beta_k x_{ik}$ → $g(\mu_i) = x_i^\beta$

CONDIZIONI RESTRITTIVE → supporto di Y non sempre è tutto \mathbb{R}
(es. va discute, o non negative)
→ distribuzione Normale non adeguata:

{Normale
Poisson
gamma
binomiale
negativa} } famiglia
esponentiale
Tale
lineare

5.1 Introduzione

Il modello di regressione lineare normale è adatto per problemi nei quali alle variabili risposta si possa assegnare una distribuzione normale, con speranza matematica combinazione lineare delle determinazioni delle variabili esplicative e con varianza costante.

In molte situazioni le precedenti condizioni sono troppo restrittive. In particolare, nelle applicazioni alla tariffazione i numeri aleatori di interesse possono essere discreti, come il numero dei sinistri che colpiscono un rischio assicurato, oppure avere come supporto la semiretta positiva e distribuzione con asimmetria positiva, come l'importo del danno provocato da un sinistro. In tali casi l'ipotesi di distribuzione normale delle variabili risposta chiaramente non è adeguata. Al numero dei sinistri è usuale assegnare distribuzione di Poisson o binomiale negativa, mentre per l'importo del danno sono frequentemente adottate le distribuzioni gamma o gaussiana inversa. Inoltre, spesso non è accettabile l'ipotesi che il legame tra la speranza matematica delle variabili risposta e le determinazioni delle variabili esplicative sia lineare. Per esempio, non rispecchiano tale struttura i modelli moltiplicativi per la tariffazione illustrati nel § 2.3. Si noti poi che per tutte le distribuzioni sopra citate sussiste un legame funzionale tra speranza matematica e varianza, per esempio, nella distribuzione di Poisson la speranza matematica è uguale alla varianza. Ciò implica che, se la speranza matematica è funzione delle variabili tarifarie, tale è anche la varianza che pertanto non può essere costante.

I modelli lineari generalizzati (*Generalized Linear Models*, GLM) estendono in due direzioni i modelli di regressione lineare. Da un lato consentono di assegnare alle variabili risposta distribuzioni appartenenti alla classe esponenziale lineare che comprende, oltre alla normale, diverse altre distribuzioni, per esempio le distribuzioni di Poisson, binomiale negativa, gamma e gaussiana inversa. Dall'altro permettono di svincolarsi dall'ipotesi

che le speranze matematiche delle variabili risposta abbiano una struttura lineare. Infatti, le speranze matematiche sono messe in relazione con i rispettivi previsori lineari mediante una funzione di collegamento che può essere diversa dalla funzione identica.

I GLM sono stati introdotti da Nelder e Wedderburn (1972) e sono stati successivamente estesi e sviluppati. Costituiscono un'ampia classe di modelli che trova largo impiego in diversi settori, ciò è dovuto sia alla loro flessibilità sia al fatto che molti pacchetti statistici, per esempio GLIM, SAS, S-PLUS, R, includono procedure che permettono di ottenere le stime dei parametri, di valutare la bontà di adattamento ai valori osservati, di confrontare e selezionare modelli.

Questo capitolo è dedicato ad una introduzione ai GLM. Sono dapprima richiamate le distribuzioni delle famiglie esponenziali lineari ed alcune loro proprietà; quindi è presentata la struttura dei GLM ed è descritta la metodologia di stima dei parametri. Alle analisi inferenziali ed alla selezione delle variabili è invece dedicato il prossimo Capitolo 6. Gli esempi e le applicazioni numeriche sono sviluppati con la procedura genmod di SAS.

Avvertiamo che la trattazione è rivolta alla presentazione degli aspetti più rilevanti, con particolare attenzione alla descrizione dei modelli per le applicazioni attuariali. Le metodologie statistiche sono presentate da un punto di vista applicativo. Per uno studio più dettagliato ed approfondito rimandiamo ai lavori sui GLM riportati in bibliografia, a loro volta ricchi di riferimenti bibliografici. Per la stesura di questo e del prossimo capitolo ci siamo basati, in particolare, su Fahrmeir, Tutz (2001) e su McCullagh, Nelder (1989).

5.2 Distribuzioni delle famiglie esponenziali lineari

In questo paragrafo presentiamo le distribuzioni delle famiglie esponenziali lineari ed alcune loro proprietà. Segnaliamo che ci limitiamo a considerare la sottoclasse delle famiglie che intervengono nei modelli lineari generalizzati unidimensionali, dette anche *famiglie esponenziali di dispersione di ordine uno*. Per approfondimenti si vedano, per esempio, Jorgensen (1997), Pace, Salvan (1996). Alcuni complementi sono riportati nell'Appendice B.

Una famiglia esponenziale lineare \mathcal{F} è una famiglia parametrica di distribuzioni di probabilità, non degeneri³, con funzione di densità (funzione di probabilità, nel caso discreto) che può essere scritta nella forma

$$f(y; \theta, \lambda) = \exp\left\{\frac{y\theta - b(\theta)}{\lambda}\right\} c(y, \lambda), \quad y \in \mathcal{Y} \subset \mathbb{R}, \quad (5.2.1)$$

dove θ e λ sono due parametri reali, $\theta \in \Theta \subset \mathbb{R}$, $\lambda \in \Lambda \subset]0, +\infty]$, b e c sono funzioni reali, $c(y, \lambda) \geq 0$. Inoltre, Θ è un intervallo non degenere, cioè tale che l'insieme dei suoi punti interni, $\text{int } \Theta$, sia non vuoto. Gli elementi che caratterizzano una famiglia esponenziale lineare \mathcal{F} sono pertanto gli insiemi Θ , Λ e le funzioni b , c .

³ Ricordiamo che si dice *supporto* di una distribuzione di probabilità su \mathbb{R} l'insieme dei numeri reali tali che la distribuzione assegna probabilità positiva ad ogni loro intorno. Una distribuzione è detta *non degenera* se il suo supporto non è costituito da un unico numero reale. Se la distribuzione è dotata di varianza, ciò equivale a dire che la distribuzione ha varianza positiva.

Si prova che b , detta *funzione cumulante*, è dotata delle derivate di ogni ordine in $\text{int}\Theta$. Si prova inoltre che la funzione cumulante caratterizza una particolare famiglia nell'ambito della classe delle famiglie esponenziali lineari, nel senso che se di una famiglia è assegnata la funzione cumulante b definita in un intervallo Θ , non degenere, allora rimangono determinati sia un insieme Λ sia la funzione c .

Il parametro θ è detto *parametro canonico* ed è collegato con la speranza matematica della distribuzione (v. in seguito (5.2.4)). Il parametro λ è detto *parametro di dispersione*. Si noti che nell'espressione della $f(y; \theta, \lambda)$ solo il primo fattore dipende dal parametro canonico θ .

In particolare, appartengono alla classe delle famiglie esponenziali lineari le famiglie normale, di Poisson, binomiale, binomiale negativa, gamma e gaussiana inversa. Gli elementi che caratterizzano tali famiglie sono riportati nella Tabella 5.1, dove μ indica la speranza matematica della distribuzione.

Segnaliamo che le distribuzioni di Poisson costituiscono una sottofamiglia di una famiglia esponenziale lineare più ampia nella quale $\Lambda =]0, +\infty[$; le distribuzioni di Poisson si ottengono fissando $\lambda = 1$. Analogamente, le distribuzioni binomiali scalate con n fissato, appartenente all'insieme \mathcal{N} dei numeri naturali, costituiscono una sottofamiglia di una famiglia esponenziale lineare più ampia nella quale $\Lambda = \{1/n, n \in \mathcal{N}\}$.

Osserviamo che non è sempre facile determinare, a partire dalle usuali parametrizzazioni, la trasformazione che consente di rappresentare le distribuzioni delle famiglie della Tabella 5.1 secondo la (5.2.1). Si tratta di individuare il parametro canonico e la funzione cumulante. Il vantaggio di rappresentare le precedenti famiglie di distribuzioni come membri della classe esponenziale lineare consiste nel fatto che, come sarà illustrato in seguito, per tali famiglie è possibile seguire una procedura generale per la stima dei parametri e per le analisi inferenziali in modelli parametrici di regressione: modelli diversi possono essere visti come membri di un'unica classe e possono essere trattati mediante un comune approccio.

Tabella 5.1. Famiglie della classe esponenziale lineare

Famiglia di distribuzioni	θ Θ	λ Λ	$b(\theta)$	$c(y, \lambda)$
Normale $N(\mu, \sigma^2)$ $\mu \in \mathbb{R}, \sigma > 0$	μ \mathbb{R}	σ^2 $]0, +\infty[$	$\theta^2/2$	$(2\pi\lambda)^{-1/2} \exp\{-y^2/(2\lambda)\}$
Poisson $P(\mu)$ $\mu > 0$	$\log \mu$ \mathbb{R}	1 $\{1\}$	$\exp(\theta)$	$1/y!$
Binomiale scalata $B(n, \pi)/n$ $n \in \mathcal{N}$ fissato, $\pi \in]0, 1[$	$\log(\pi/(1-\pi))$ \mathbb{R}	$1/n$ $\{1/n\}$	$\log(1+e^\theta)$	$\binom{1/\lambda}{y/\lambda}$
Binomiale negativa $BN(\mu, \alpha)$ $\alpha > 0$ fissato, $\mu > 0$	$\log(\mu/(\alpha+\mu))$ $]-\infty, 0[$	1 $\{1\}$	$-\alpha \log(1-e^\theta)$	$\Gamma(\alpha+y)/(\Gamma(\alpha)y!)$
Gamma $G(\alpha, \mu)$ $\alpha > 0, \mu > 0$	$-1/\mu$ $]-\infty, 0[$	$1/\alpha$ $]0, +\infty[$	$-\log(-\theta)$	$(1/\lambda)^{1/\lambda} y^{1/\lambda - 1}/\Gamma(1/\lambda)$
Gaussiana inversa $GI(\mu, \beta)$ $\mu > 0, \beta > 0$	$-(2\mu^2)^{-1}$ $]-\infty, 0[$	β $]0, +\infty[$	$-(-2\theta)^{1/2}$	$(2\pi\lambda y^3)^{-1/2} \exp\{-1/(2\lambda y)\}$

Distribuzioni con peso assegnato

In molti problemi, è naturale considerare distribuzioni di una famiglia esponenziale lineare in cui il parametro di dispersione è del tipo $\lambda = \phi/\omega$, dove $\omega > 0$ è un peso assegnato e $\phi > 0$ è un parametro che continuiamo a chiamare *parametro di dispersione*. La funzione di probabilità o di densità dipende dunque dai due parametri θ e ϕ ed ha un'espressione del tipo

$$f(y; \theta, \phi, \omega) = \exp \left\{ \frac{\omega}{\phi} [y\theta - b(\theta)] \right\} c(y, \phi, \omega), \quad y \in \mathcal{Y} \subset \mathbb{R}, \quad (5.2.2)$$

per $\theta \in \Theta \subset \mathbb{R}$ e ϕ tale che $\phi/\omega \in \Lambda \subset]0, +\infty[$. In particolare, se $\omega = 1$, $\phi = \lambda \in \Lambda$. Le distribuzioni delle famiglie della Tabella 5.1 possono allora essere viste come distribuzioni con peso unitario.

Si noti che, per distribuzioni appartenenti a famiglie esponenziali lineari con $\Lambda =]0, +\infty[$, fissato un peso positivo arbitrario, ϕ può essere un qualunque numero positivo. In particolare, ciò accade per le famiglie normale, gamma, gaussiana inversa. Le distribuzioni appartenenti alla famiglia normale, con funzione di densità (5.2.2) e $\omega > 0$ fissato sono dette anche *distribuzioni normali con peso ω* . Analoga terminologia è utilizzata per le distribuzioni con peso delle altre due famiglie. Le distribuzioni del tipo (5.2.2) con funzione cumulante $b(\theta) = \exp(\theta)$, $\theta \in \mathbb{R}$, $\omega > 0$ fissato e $\phi = 1$, appartengono alla famiglia esponenziale lineare che contiene le distribuzioni di Poisson: tali distribuzioni sono dette anche *distribuzioni di Poisson con peso ω* . Le distribuzioni con funzione cumulante $b(\theta) = \log(1 + e^\theta)$, $\theta \in \mathbb{R}$, $\omega \in \mathcal{W}$ fissato e $\phi = 1$, che appartengono alla famiglia esponenziale lineare che contiene le distribuzioni binomiali scalate, sono dette anche *distribuzioni binomiali con peso ω* .

Vedremo in seguito, nei modelli per le applicazioni attuariali, diversi esempi di distribuzioni con peso.

Funzione generatrice dei momenti e momenti

Le distribuzioni delle famiglie esponenziali lineari sono dotate di funzione generatrice dei momenti⁴. Sia Y un numero aleatorio con distribuzione appartenente ad una famiglia esponenziale lineare \mathcal{F} , con densità $f(y; \theta, \lambda)$ data dalla (5.2.1), con $\theta \in \text{int } \Theta$, $\lambda \in \Lambda$, la funzione generatrice dei momenti di Y è

$$m_Y(t; \theta, \lambda) = \exp \left\{ \frac{b(\theta + t\lambda) - b(\theta)}{\lambda} \right\}, \quad (5.2.3)$$

per ogni t tale che $\theta + t\lambda \in \Theta$.

La distribuzione ha momenti finiti di ogni ordine e si ha

$$E(Y^n) = \frac{d^n}{dt^n} m_Y(t; \theta, \lambda) \Big|_{t=0}.$$

⁴ Ricordiamo che una distribuzione di probabilità (un numero aleatorio Y), con funzione di ripartizione F , si dice *dotata di funzione generatrice dei momenti* se esiste un intorno I_0 di zero tale che, per ogni $t \in I_0$, l'integrale $\int_{\mathcal{Y}} e^{ty} dF(y)$ è finito. In tale caso, la funzione $m(t) = \int_{\mathcal{Y}} e^{ty} dF(y)$, $t \in I_0$, è detta *funzione generatrice dei momenti della distribuzione*. Una distribuzione dotata di funzione generatrice dei momenti ha momenti finiti di ogni ordine; il momento n -esimo è la derivata n -esima della funzione generatrice dei momenti in $t=0$. Inoltre, la funzione generatrice dei momenti determina univocamente la distribuzione.

In particolare, si verifica facilmente che la speranza matematica e la varianza sono date dalle

$$E(Y) = \mu = b'(\theta) \quad (5.2.4)$$

e

$$\text{var}(Y) = \lambda b''(\theta). \quad \Rightarrow b''(\theta) > 0 \quad \forall \theta \in \Theta \quad (5.2.5)$$

Si osservi che per le distribuzioni di una famiglia esponenziale lineare le speranze matematiche non dipendono dal parametro di dispersione. Fissato θ , le varianze differiscono invece per un fattore di scala, che è pari proprio al parametro di dispersione. I parametri θ e λ racchiudono dunque, in particolare, le informazioni sulla speranza matematica e sulla varianza della distribuzione.

Se la distribuzione di Y è del tipo (5.2.2), con $\theta \in \text{int}\Theta$, la funzione generatrice dei momenti è

$$m_Y(t; \theta, \phi, \omega) = \exp \left\{ \frac{\omega}{\phi} [b(\theta + t\phi/\omega) - b(\theta)] \right\}; \quad (5.2.6)$$

la varianza della distribuzione è

$$\text{var}(Y) = \frac{\phi}{\omega} b''(\theta). \quad (5.2.7)$$

► Funzione di varianza

Assegnata una famiglia esponenziale lineare \mathcal{F} , si prova che la derivata prima b' della funzione cumulante, definita in $\text{int}\Theta$, è crescente e pertanto $b': \text{int}\Theta \rightarrow M$, con $M = b'(\text{int}\Theta)$, è invertibile. L'insieme $M = b'(\text{int}\Theta)$ è detto anche *spazio dei valori attesi* e rappresenta l'insieme dei valori ammissibili per la speranza matematica delle distribuzioni della famiglia, per $\theta \in \text{int}\Theta$.

La funzione

$$V(\mu) = b''(b'^{-1}(\mu)), \quad \mu \in M, \quad V: M \rightarrow \mathbb{R}^+$$

è detta *funzione di varianza*. La varianza di Y può allora essere espressa attraverso il parametro μ e si ha

$$\text{var}(Y) = \lambda V(\mu) = \lambda b''(b'^{-1}(\mu))$$

o, nel caso delle distribuzioni del tipo (5.2.2),

$$\text{var}(Y) = \frac{\phi}{\omega} V(\mu).$$

Nella Tabella 5.2 sono riportate le funzioni di varianza per le famiglie di distribuzioni della Tabella 5.1.

Si può provare che la funzione di varianza $V(\mu)$, $\mu \in M$, caratterizza una particolare famiglia nella classe delle famiglie esponenziali lineari, nel senso che la funzione di varianza determina la funzione cumulante b e l'insieme dei parametri canonici Θ . Tale proprietà ha una portata operativa in quanto consente di assegnare una famiglia esponenziale lineare attraverso la funzione di varianza anziché la funzione cumulante.

Tabella 5.2. Funzioni di varianza

Famiglia di distribuzioni	$V(\mu)$	M
Normale	1	\mathfrak{N}
Poisson	μ	$]0, +\infty[$
Binomiale scalata	$\mu(1-\mu)$	$]0, 1[$
Binomiale negativa	$\mu(1+\mu/\alpha)$	$]0, +\infty[$
Gamma	μ^2	$]0, +\infty[$
Gaussiana inversa	μ^3	$]0, +\infty[$

► Funzioni di varianza di tipo potenza

Una importante classe di famiglie esponenziali lineari è rappresentata dalle famiglie con funzioni di varianza di tipo potenza

$$V_\xi(\mu) = \mu^\xi, \quad \mu \in M_\xi = \begin{cases}]0, +\infty[& \text{se } \xi \neq 0 \\ \mathfrak{N} & \text{se } \xi = 0. \end{cases}$$

I modelli corrispondenti sono anche detti *modelli di Tweedie*. Si tratta di una classe che comprende, rispettivamente per $\xi = 0, 1, 2, 3$, le famiglie normale, di Poisson, gamma e gaussiana inversa. Per $1 < \xi < 2$ si ottengono famiglie di distribuzioni di tipo Poisson-composto (v. § 9.2), per $\xi \geq 2$ famiglie di distribuzioni dotate di densità con supporto l'intervallo $]0, +\infty[$. Non esistono invece famiglie esponenziali lineari con funzioni di varianza di tipo potenza per $0 < \xi < 1$.

5.3 La classe dei modelli lineari generalizzati

I GLM sono modelli di regressione che generalizzano i lineari. Con riferimento a n unità statistiche, si dispone di un insieme di osservazioni $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$, dove per ciascuna unità statistica, quindi per $i = 1, \dots, n$, y_i rappresenta il valore di una grandezza di interesse e \mathbf{x}_i il vettore delle determinazioni assunte da un insieme di variabili esplicative. Il vettore $\mathbf{y} = (y_1, \dots, y_n)'$ è visto come valore osservato del vettore aleatorio $\mathbf{Y} = (Y_1, \dots, Y_n)'$ delle variabili risposta, per il quale è formulata una ipotesi probabilistica che mette in relazione la distribuzione di \mathbf{Y} con i vettori delle determinazioni delle variabili esplicative.

In sintesi, un GLM è definito dalle seguenti ipotesi.

- Ipotesi probabilistiche. Le variabili risposta Y_1, \dots, Y_n sono stocasticamente indipendenti, con distribuzioni appartenenti ad una medesima famiglia esponenziale lineare.
- Ipotesi strutturali. Vi è un legame tra la speranza matematica μ_i della variabile risposta Y_i ed il vettore \mathbf{x}_i delle determinazioni delle variabili esplicative, relativi all' i -esima osservazione, espresso dalla

$$g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta},$$

dove $\boldsymbol{\beta}$ è un vettore di parametri e g una funzione di collegamento, invertibile. Si ha pertanto

$$E(Y_i) = \mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}).$$

Descriviamo ora in dettaglio i diversi elementi che intervengono nella definizione di un GLM.

► Distribuzione del vettore delle variabili risposta

Le variabili risposta Y_1, \dots, Y_n sono stocasticamente indipendenti, con distribuzioni appartenenti ad una medesima famiglia esponenziale lineare. La funzione di probabilità o di densità di Y_i è del tipo

$$f(y; \theta_i, \phi, \omega_i) = \exp\left\{\frac{\omega_i}{\phi}[y\theta_i - b(\theta_i)]\right\} c(y, \phi, \omega_i), \quad (5.3.1)$$

dove θ_i e ϕ sono i parametri, canonico e di dispersione, e $\omega_i > 0$ è un peso assegnato. Inoltre, si assume che il supporto delle distribuzioni di Y_1, \dots, Y_n non dipenda dai parametri $\theta_1, \dots, \theta_n, \phi$.⁵

Si noti che, essendo fissata la famiglia esponenziale lineare, la funzione cumulante b non varia al variare di i . Inoltre, si suppone che non dipenda da i il parametro ϕ di dispersione, mentre, in generale, dipendono da i il parametro canonico θ_i e il peso ω_i .

Per quanto visto nel paragrafo precedente in relazione ai momenti delle distribuzioni della classe esponenziale lineare, si ha

$$E(Y_i) = \mu_i = b'(\theta_i)$$

e

$$\text{var}(Y_i) = \frac{\phi}{\omega_i} b''(\theta_i) = \frac{\phi}{\omega_i} V(\mu_i),$$

dove $V(\mu) = b''(b'^{-1}(\mu))$ è la funzione di varianza della famiglia delle distribuzioni assegnate alle variabili risposta.

Osserviamo che l'ipotesi di invarianza del parametro ϕ di dispersione rispetto ad i comporta, per esempio, che se le variabili risposta hanno distribuzioni normali con lo stesso peso, allora hanno tutte la stessa varianza; se le variabili risposta hanno distribuzioni gamma con lo stesso peso, allora hanno tutte lo stesso coefficiente di variazione.

Con riferimento ai pesi, si noti che, a parità di ϕ e $V(\mu_i)$, la varianza di Y_i è tanto maggiore quanto minore è il peso ω_i . I pesi possono allora essere utilizzati per incorporare nel modello informazioni sull'affidabilità delle singole osservazioni.

► Le variabili esplicative

Ad ognuna delle n unità statistiche è associato un vettore di determinazioni di variabili esplicative che rappresentano caratteristiche osservabili, influenti sulle valutazioni probabilistiche delle variabili risposta.

⁵ Tale ipotesi è introdotta per motivi tecnici legati al problema di determinare le stime dei parametri che, come sarà chiarito nel prossimo paragrafo, si ottengono con il metodo della massima verosimiglianza, e affinché possano sussistere le proprietà asintotiche dei relativi stimatori.

Le variabili di classificazione sono codificate con variabili indicatrici nel modo indicato nel § 3.3. La descrizione delle caratteristiche avviene quindi con m variabili numeriche, X_1, \dots, X_m . Siano x_{i1}, \dots, x_{im} le determinazioni di tali variabili per l' i -esima osservazione.

Indichiamo con \mathbf{X} la matrice in cui la prima colonna è tutta di elementi unitari e la $(j+1)$ -esima colonna riporta le determinazioni x_{ij} della variabile X_j per ogni osservazione i

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \vdots & \vdots & & & & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \vdots & \vdots & & & & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix}.$$

La i -esima riga di \mathbf{X} riporta quindi le determinazioni di tutte le variabili esplicative per l' i -esima osservazione, con l'aggiunta di $x_{i0} = 1$. Poniamo $p = m+1$.

La matrice \mathbf{X} , di tipo $n \times p$, è detta *matrice di regressione* o *matrice delle condizioni sperimentali (design matrix)*. Nel seguito supponiamo che riesca $n > p$ e che \mathbf{X} sia di rango pieno p . Le p colonne della matrice sono dunque linearmente indipendenti⁶.

Sottolineiamo che le variabili esplicative sono trattate come elementi non aleatori.

► Il previsore lineare

Il vettore delle determinazioni delle variabili esplicative $\mathbf{x}'_i = (1, x_{i1}, \dots, x_{im})$ influisce sulla distribuzione della variabile risposta Y_i tramite il *previsore lineare relativo all' i -esima osservazione*

$$\eta_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{im}\beta_m = \mathbf{x}'_i \boldsymbol{\beta},$$

dove $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)' \in \mathcal{B} \subset \Re^p$ è un vettore di parametri, comuni a tutte le unità statistiche.

Il previsore lineare $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ rappresenta la *componente sistematica* del modello ed è funzione lineare dei parametri $\beta_0, \beta_1, \dots, \beta_m$.

I parametri di regressione $\beta_0, \beta_1, \dots, \beta_m$ sono considerati certi, ma non noti. Il parametro β_0 è detto *intercetta* e può non essere presente nel modello: $\beta_0 = 0$. In tale caso non compaiono in \mathbf{X} la prima colonna di termini unitari e nell'espressione di η_i l'addendo β_0 . Nella descrizione che segue supponiamo che il modello contenga l'intercetta.

► La funzione di collegamento

È una funzione g reale di variabile reale, invertibile, detta anche *funzione di legame (link function)*, che mette in relazione le componenti del previsore lineare con le *speranze matematiche* delle variabili risposta. Si ha

⁶ Le variabili esplicative del modello le cui determinazioni costituiscono la matrice di regressione \mathbf{X} sono, in generale, funzioni delle variabili esplicative "originali". Infatti, (v. Capitolo 3 e § 5.6) possono essere trasformate di variabili quantitative oppure variabili indicatrici introdotte per codificare variabili di classificazione o ancora variabili che codificano interazioni. Sarebbe dunque opportuno distinguere, anche usando simboli diversi, le variabili della matrice di regressione dalle variabili esplicative "originali". Nei successivi paragrafi, per non introdurre nuovi simboli, non seguiremo questa strada: la natura delle variabili apparirà chiara dal contesto.

$$\eta_i = g(\mu_i), \quad i = 1, \dots, n.$$

Ne segue che

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(x'_i \beta). \quad (5.3.2)$$

Osserviamo che il dominio della funzione g deve essere lo spazio dei valori attesi, $M = b'(\text{int } \Theta)$, della famiglia di distribuzioni delle variabili risposta. Per quanto riguarda l'insieme immagine $g(M)$, è opportuno che questo coincida con \mathfrak{N} . In tale caso, qualunque sia $\beta \in \mathfrak{N}^p$, $x'_i \beta \in g(M)$ ovvero $g^{-1}(x'_i \beta)$ è un valore ammissibile per μ_i . Si può allora assumere $\mathcal{B} = \mathfrak{N}^p$ come spazio parametrico per β . Se invece l'insieme immagine di g non è \mathfrak{N} , lo spazio parametrico \mathcal{B} deve essere tale che per ogni $\beta \in \mathcal{B}$, $x'_i \beta \in g(M)$. In generale, ciò comporta che si devono considerare vincoli per β , che non può pertanto variare in tutto \mathfrak{N}^p .

Per motivi tecnici, legati al problema di determinare le stime dei parametri, la funzione di collegamento deve soddisfare opportune condizioni di regolarità. Pertanto, richiediamo fin d'ora che g sia strettamente monotona e che ammetta derivate prima e seconda continue. (invertibilità)

Se è dato il vettore dei parametri β , tramite la (5.3.2) la funzione di collegamento consente di ottenere le speranze matematiche delle variabili risposta a partire dalle determinazioni delle variabili esplicative. In particolare, nella tariffazione la funzione di collegamento indica, per esempio, come calcolare il premio equo in funzione delle caratteristiche tariffarie, quindi attraverso la funzione g si determina il modello tariffario.

Se g è la funzione identica, si ha $\mu_i = x'_i \beta = \sum_{j=0}^m x_{ij} \beta_j$ e si ottiene il modello tariffario additivo (v. in seguito § 5.6). Come osservato sopra, a meno di non considerare vincoli per β , il previsore lineare può assumere valori in tutto \mathfrak{N} ; pertanto la funzione di collegamento identica può essere adeguata per distribuzioni delle variabili risposta per le quali la speranza matematica possa assumere qualunque valore reale, quali le normali. Con riferimento a problemi di conteggio in cui le variabili risposta hanno speranze matematiche positive, tale funzione di collegamento può invece non essere opportuna.

Se g è la funzione logaritmo, si ha $\mu_i = e^{x'_i \beta} = \prod_{j=0}^m e^{x_{ij} \beta_j}$ e si ottiene il modello tariffario moltiplicativo (v. in seguito § 5.6). Con questa funzione di collegamento, gli effetti additivi del previsore lineare si trasformano in effetti moltiplicativi sul premio equo. Inoltre, μ_i è sempre positiva.

Una classe di funzioni di collegamento dipendente da un parametro reale γ è costituita dalla classe delle funzioni potenza

$$g(\mu) = \begin{cases} \frac{\mu^\gamma - 1}{\gamma} & \text{se } \gamma \neq 0 \\ \log \mu & \text{se } \gamma = 0. \end{cases} \quad (5.3.3)$$

Per $\gamma = 1$ si ottiene $g(\mu) = \mu - 1$ e quindi, a meno di una traslazione, la funzione di collegamento identica. Per $\gamma \rightarrow 0$, si ha che $g(\mu) \rightarrow \log \mu$. Al variare di γ in $[0, 1]$, la g definita dalla (5.3.3) consente di passare "con continuità" dal modello additivo al modello moltiplicativo.

Una specificazione alternativa della classe delle funzioni di collegamento di tipo potenza è la seguente

$$g(\mu) = \begin{cases} \mu^\gamma & \text{se } \gamma \neq 0 \\ \log \mu & \text{se } \gamma = 0. \end{cases} \quad (5.3.4)$$

In ogni famiglia esponenziale lineare, la funzione b'^{-1} trasforma la speranza matematica μ nel parametro canonico θ . Infatti, da $\mu = b'(\theta)$ e dall'invertibilità di b' , segue $b'^{-1}(\mu) = \theta$. Scegliendo $g(\mu) = b'^{-1}(\mu)$ come funzione di collegamento in un GLM, si ha

$$\eta_i = g(\mu_i) = \theta_i, \quad i = 1, \dots, n.$$

Tale funzione è detta *funzione canonica di collegamento* perché mette direttamente in collegamento il previsore lineare con il parametro canonico che è dunque espresso come combinazione lineare delle determinazioni delle variabili esplicative.

Per *g* funzione canonica di collegamento si ha

$$g'(\mu) = \frac{1}{b''(b'^{-1}(\mu))} = \frac{1}{V(\mu)}. \quad (5.3.5)$$

Nell'ultima colonna della Tabella 5.3 sono riportate le funzioni canoniche di collegamento per le famiglie di distribuzioni della Tabella 5.1.

In particolare, per la famiglia normale il collegamento canonico è la funzione identica, per la famiglia di Poisson è la funzione logaritmo, per la binomiale scalata è la funzione *logit*, per la gamma è l'opposta della reciproca.

I GLM con funzione canonica di collegamento hanno importanti proprietà statistiche ed inoltre conducono a semplificazioni matematiche nel procedimento di stima dei parametri (v. successivo § 5.5). Tuttavia, in alcuni casi, funzioni di collegamento diverse dalla canonica possono risultare più adeguate. Per esempio, per le distribuzioni della famiglia gamma, il collegamento canonico comporta $\mu_i = -1/\eta_i$. Se non si pongono vincoli a β , $-1/\eta_i$ potrebbe risultare negativo, mentre la speranza matematica delle distribuzioni gamma è positiva. Può allora essere preferibile considerare come funzione di collegamento, per esempio, il logaritmo. La scelta della funzione di collegamento dipende dunque dalla famiglia esponenziale lineare considerata per le distribuzioni delle variabili risposta ed anche dal problema trattato.

Riassumendo, un GLM è caratterizzato da tre componenti: la famiglia esponenziale lineare delle distribuzioni assegnate alle variabili risposta, la matrice delle determinazioni delle variabili esplicative con l'associato previsore lineare, la funzione di collegamento.

Osserviamo che, in forza del legame che sussiste tra la speranza matematica e la varianza per le distribuzioni delle famiglie esponenziali lineari, la specificazione di una particolare struttura per la speranza matematica di Y_i , $\mu_i = g^{-1}(x'_i \beta)$, induce una struttura per la varianza; si ha infatti

$$var(Y_i) = \frac{\phi}{\omega_i} V(g^{-1}(x'_i \beta)).$$

La varianza dipende dunque dalla stessa combinazione lineare delle variabili esplicative da cui dipende $E(Y_i)$.

Tabella 5.3. Funzioni canoniche di collegamento

Famiglia di distribuzioni	Funzione cumulante $b(\theta)$	Derivata $b'(\theta)$	Collegamento canonico $g(\mu) = b'^{-1}(\mu)$
Normale	$\theta^2 / 2$	θ	μ
Poisson	$\exp(\theta)$	$\exp(\theta)$	$\log \mu$
Binomiale scalata	$\log(1 + e^\theta)$	$\frac{\exp(\theta)}{1 + \exp(\theta)}$	$\log(\mu/(1-\mu))$
Binomiale negativa	$-\alpha \log(1 - e^\theta)$	$\alpha \frac{\exp(\theta)}{1 - \exp(\theta)}$	$\log(\mu/(\alpha+\mu))$
Gamma	$-\log(-\theta)$	$-1/\theta$	$-1/\mu$
Gaussiana Inversa	$-(2\theta)^{1/2}$	$(-2\theta)^{-1/2}$	$-1/(2\mu^2)$

I parametri

In un GLM intervengono i parametri canonici $\theta_1, \dots, \theta_n$ ed il parametro di dispersione ϕ . Essi sono considerati certi, ma non noti e si stimano dai dati $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$. In alcuni casi ϕ è noto: per esempio, per le distribuzioni di Poisson, si ha $\phi = 1$.

Usualmente, nei GLM, la stima dei parametri canonici $\theta_1, \dots, \theta_n$ avviene stimando il vettore $\boldsymbol{\beta}$ dei parametri di regressione. Dato $\boldsymbol{\beta}$, rimangono determinati i parametri canonici. Infatti, assegnata la matrice \mathbf{X} , il vettore $\boldsymbol{\beta}$ consente di determinare η_i tramite la $\eta_i = \mathbf{x}'_i \boldsymbol{\beta}$. A sua volta η_i , data la funzione di collegamento g , individua μ_i tramite la $\mu_i = g^{-1}(\eta_i)$. Infine, μ_i individua θ_i tramite la $\theta_i = b'^{-1}(\mu_i)$. Si hanno cioè le

$$\theta_i = b'^{-1}(g^{-1}(\mathbf{x}'_i \boldsymbol{\beta})), \quad i = 1, \dots, n, \quad (5.3.6)$$

che mettono in relazione i parametri di regressione con i parametri canonici.

In particolare, se g è il collegamento canonico, si ha

$$\theta_i = \eta_i = \mathbf{x}'_i \boldsymbol{\beta}, \quad i = 1, \dots, n. \quad (5.3.7)$$

Dalle precedenti relazioni appare evidente che, per unità statistiche con un medesimo vettore di determinazioni delle variabili esplicative, si ottiene lo stesso valore del parametro canonico e pertanto, a meno dei pesi, le medesime distribuzioni delle corrispondenti variabili risposta.

Concludiamo il paragrafo osservando che i GLM costituiscono una classe ampia e flessibile di modelli che consentono di considerare diverse famiglie di distribuzioni per le variabili risposta e di tenere conto, mediante l'introduzione di opportune variabili esplicative, di fattori che si ritiene abbiano influenza significativa sulle distribuzioni da stimare. Inoltre, molti usuali modelli di regressione possono essere rivisti in tale ambito. Rientrano, infatti, nella classe dei GLM i modelli di regressione lineare con variabili risposta con distribuzione normale, i modelli log-lineari, i modelli di regressione logistica e di Poisson. Sottolineiamo tuttavia che anche i GLM presentano ipotesi che li rendono inadatti per trattare alcuni problemi: l'indipendenza delle variabili risposta, la richiesta che le distribuzioni siano completamente specificate e appartengano ad una famiglia esponenziale lineare, il fatto che le speranze matematiche e le varianze di tali distribuzioni dipendano dalle

medesime variabili esplicative, l'ipotesi che il parametro di dispersione sia comune a tutte le variabili risposta. In letteratura sono proposte estensioni dei GLM volte a superare i limiti sopra evidenziati. Nel § 5.9 è presentata una di tali estensioni, relativa ai modelli con quasi-verosimiglianza, che allenta l'ipotesi sulle distribuzioni delle variabili risposta.

5.4 La stima dei parametri

Come detto nel § 5.3, la stima dei parametri canonici nell'ambito dei GLM si ottiene dalla stima dei parametri di regressione. Per il parametro vettoriale β si ricorre usualmente al metodo della massima verosimiglianza. La scelta è motivata dalle proprietà dei corrispondenti estimatori, dalle quali discendono alcuni risultati sulle distribuzioni delle statistiche utilizzate per l'inferenza. Alcune di tali proprietà sono richiamate sinteticamente nell'Appendice C. Nel seguito supponiamo soddisfatte le condizioni di regolarità che consentono di effettuare i passaggi via via indicati⁷.

Sia $y = (y_1, \dots, y_n)'$ il valore osservato del vettore delle variabili risposta $Y = (Y_1, \dots, Y_n)'$. La log-verosimiglianza l dell'osservazione y come funzione dei parametri $\theta = (\theta_1, \dots, \theta_n)$ e ϕ è

$$l(\theta, \phi; y) = \log L(\theta, \phi; y) = \sum_{i=1}^n \left\{ \frac{\omega_i}{\phi} [y_i \theta_i - b(\theta_i)] + \log c(y_i, \phi, \omega_i) \right\} = \sum_{i=1}^n l_i(\theta_i, \phi; y_i).$$

Dapprima, supponiamo che il parametro di dispersione ϕ sia fissato. Per i modelli per i quali ϕ non è dato, ciò equivale a considerare una restrizione della log-verosimiglianza. Si vedrà che, ai fini di ottenere la stima di β , la condizione non è restrittiva.

Per la (5.3.6), ciascun parametro canonico θ_i , $i = 1, \dots, n$, può essere visto come funzione di β . Indichiamo con il simbolo $\ell(\beta) = \sum_{i=1}^n \ell_i(\beta)$ la log-verosimiglianza come funzione composta di β .

Generalmente, le stime di massima verosimiglianza sono determinate come le soluzioni del sistema delle equazioni di verosimiglianza,

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i(\beta)}{\partial \beta_j} = 0, \quad j = 0, \dots, m,$$

in corrispondenza delle quali la matrice hessiana $\left[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_h} \right]_{j,h}$ risulta definita negativa.

Si individuano così i punti di massimo relativo di ℓ . Osserviamo che, se la funzione di log-verosimiglianza è concava, i punti di massimo relativo sono anche punti di massimo assoluto. In particolare, se la log-verosimiglianza è strettamente concava, come accade



⁷ Le condizioni di regolarità cui si fa riferimento sono le usuali ipotesi che consentono di effettuare l'analisi del secondo ordine per i punti di ottimo relativo. Oltre alle ipotesi già introdotte sul supporto delle distribuzioni delle variabili risposta (indipendenti dai parametri), sul rango della matrice del modello (rango pieno di X) e sulla funzione di collegamento g (strettamente monotona e dotata di derivate prima e seconda continue), si suppone ancora che la derivata prima di g sia ovunque non nulla, che lo spazio dei parametri di regressione \mathcal{B} sia aperto e tale che $g^{-1}(x'_i \beta) \in M$, per ogni $\beta \in \mathcal{B}$, essendo M lo spazio dei valori attesi della famiglia esponenziale lineare delle distribuzioni assegnate alle variabili risposta.

per molti importanti GLM, per esempio per ogni modello con funzione canonica di collegamento (v. più avanti, in questo paragrafo), la stima di massima verosimiglianza, se esiste, è unica. Non ci soffermiamo a discutere il problema dell'esistenza ed unicità delle stime: per qualche ulteriore commento rimandiamo alla Nota alla fine del § 5.5.

Calcoliamo ora in modo esplicito le derivate parziali della log-verosimiglianza. Schematizzando i passi che hanno condotto alla (5.3.6), si ha la seguente dipendenza di ℓ da β

$$\beta \rightarrow \eta_i = x'_i \beta \rightarrow \mu_i = g^{-1}(\eta_i) \rightarrow \theta_i = b'^{-1}(\mu_i) \rightarrow l_i(\theta_i, \phi; y_i).$$

Si ha dunque

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n \left[\frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{\partial \beta_j} \right],$$

con

$$\begin{aligned} l_i &= \frac{\omega_i}{\phi} [y_i \theta_i - l(\theta_i)] + C \\ \theta_i &= b'^{-1}(\mu_i) \quad \frac{d\theta_i}{d\mu_i} = \frac{db'^{-1}(\mu_i)}{d\mu_i} = \frac{1}{b''(b'^{-1}(\mu_i))} = \frac{1}{V(\mu_i)}, \\ \mu_i &= g^{-1}(\eta_i) \quad \frac{d\mu_i}{d\eta_i} = \frac{dg^{-1}(\eta_i)}{d\eta_i} = \frac{1}{g'(g^{-1}(\eta_i))} = \frac{1}{g'(\mu_i)}, \\ \eta_i &= x'_i \beta \quad \frac{d\eta_i}{\partial \beta_j} = x_{ij}. \end{aligned} \tag{5.4.1}$$

$$g_i = b'^{-1}(g^{-1}(x'_i \beta))$$

$\Rightarrow g(\cdot) = b'^{-1}(\cdot)$ funz. colleg canonico

$$\Rightarrow \theta_i = \eta_i = x'_i \beta$$

$$\mu = E[\eta]$$

Esprimendo le derivate parziali di ℓ tramite le componenti di $\mu = (\mu_1, \dots, \mu_n)'$, vettore delle speranze matematiche delle variabili risposta, si ottiene

$$s_j(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{1}{g'(\mu_i) V(\mu_i)}, \tag{5.4.2}$$

dove $\mu_i = g^{-1}(x'_i \beta)$.

Se la funzione di collegamento è la canonica, per la (5.3.5), riesce $g'(\mu_i) = 1/V(\mu_i)$ e quindi la (5.4.2) si semplifica nella

$$s_j(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i). \tag{5.4.3}$$

La stima di massima verosimiglianza di β si trova tra le soluzioni del sistema $s_j(\beta) = 0$, $j = 0, \dots, m$. In modo esplicito, nel caso generale, le equazioni sono

$$\sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{1}{g'(\mu_i) V(\mu_i)} = 0, \quad j = 0, \dots, m. \tag{5.4.4}$$

Si osserva immediatamente che la soluzione delle (5.4.4) non dipende da ϕ . Ciò spiega perché, ai fini della stima del vettore dei parametri di regressione β , non è restrittivo supporre ϕ noto. Si osserva inoltre che, in relazione alle distribuzioni delle variabili risposta, le equazioni di verosimiglianza dipendono solo dalle speranze matematiche e dalle varianze.

In generale, il sistema delle equazioni di verosimiglianza (5.4.4) non ammette soluzione esplicita. Per la risoluzione si ricorre a metodi numerici. Nel prossimo paragrafo

sono descritti i metodi di Newton-Raphson e *scoring* di Fisher, usualmente utilizzati nei pacchetti statistici che forniscono le stime dei parametri nei GLM.

Il vettore $s(\boldsymbol{\beta}) = (s_0(\boldsymbol{\beta}), \dots, s_m(\boldsymbol{\beta}))'$, gradiente della funzione di log-verosimiglianza, è detto *vettore di punteggio* o *vettore score*; visto come funzione di $\boldsymbol{\beta}$ è anche detto *funzione di punteggio* o *funzione score*. Indicata con V la matrice delle varianze e covarianze del vettore \mathbf{Y} ,

$$V = \text{var}(\mathbf{Y}) = \text{diag}[\text{var}(Y_i)] = \text{diag}\left[\frac{\phi}{\omega_i} V(\mu_i)\right],$$

e posto

$$\mathbf{D} = \text{diag}\left[\frac{d\mu_i}{d\eta_i}\right] = \text{diag}\left[\frac{1}{g'(\mu_i)}\right],$$

la funzione di punteggio può essere rappresentata con la seguente scrittura vettoriale

$$s(\boldsymbol{\beta}) = \mathbf{X}' \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \quad (5.4.5)$$

dove \mathbf{D} , \mathbf{V} , e $\boldsymbol{\mu}$ sono funzioni di $\boldsymbol{\beta}$.

Calcoliamo ora le derivate seconde della funzione di log-verosimiglianza. Si ha

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_h} &= \frac{\partial}{\partial \beta_h} \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} \right) = \sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} \frac{\partial}{\partial \beta_h} \left[(y_i - \mu_i) \frac{1}{g'(\mu_i) V(\mu_i)} \right] \\ &= \sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} \left[\left(\frac{\partial}{\partial \beta_h} (y_i - \mu_i) \right) \frac{1}{g'(\mu_i) V(\mu_i)} + (y_i - \mu_i) \frac{\partial}{\partial \beta_h} \left[\frac{1}{g'(\mu_i) V(\mu_i)} \right] \right]. \end{aligned}$$

Dalle terza e quarta delle (5.4.1), si ricava

$$\frac{\partial \mu_i}{\partial \beta_h} = \frac{x_{ih}}{g'(\mu_i)}.$$

Si ha quindi

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_h} = - \sum_{i=1}^n x_{ij} x_{ih} \frac{\omega_i}{\phi} \frac{1}{g'(\mu_i)^2 V(\mu_i)} + \sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{\partial}{\partial \beta_h} \left[\frac{1}{g'(\mu_i) V(\mu_i)} \right],$$

dove

$$\frac{\partial}{\partial \beta_h} \left[\frac{1}{g'(\mu_i) V(\mu_i)} \right] = -x_{ih} \frac{g''(\mu_i) V(\mu_i) + g'(\mu_i) V'(\mu_i)}{g'(\mu_i)^3 V(\mu_i)^2}.$$

→ Poniamo

$$\mathbf{H}(\boldsymbol{\beta}) = \left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_h} \right]_{j,h}, \quad (5.4.6)$$

la matrice hessiana della log-verosimiglianza, la cui opposta, $-\mathbf{H}(\boldsymbol{\beta})$, è anche detta *matrice di informazione osservata*. È facile verificare che si ha

$$-\mathbf{H}(\boldsymbol{\beta}) = \mathbf{X}' \mathbf{W}_{oss}(\boldsymbol{\beta}) \mathbf{X}, \quad (5.4.7)$$

dove

$$W_{oss}(\beta) = \text{diag} \left[\frac{\omega_i}{\phi g'(\mu_i)^2 V(\mu_i)} + \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{g''(\mu_i)V(\mu_i) + g'(\mu_i)V'(\mu_i)}{g'(\mu_i)^3 V(\mu_i)^2} \right].$$

Consideriamo ancora la matrice di informazione attesa (o matrice di informazione di Fisher)

$$\mathcal{J}(\beta) = E \left[-\frac{\partial^2 \tilde{\ell}(\beta)}{\partial \beta_j \partial \beta_h} \right]_{j,h}, \quad (5.4.8)$$

dove $\tilde{\ell}$ è il numero aleatorio che si ottiene sostituendo nell'espressione della funzione di log-verosimiglianza ℓ , i numeri aleatori Y_1, \dots, Y_n alle osservazioni y_1, \dots, y_n . L'elemento di posto (j,h) della matrice $\mathcal{J}(\beta)$ è

$$E \left(-\frac{\partial^2 \tilde{\ell}(\beta)}{\partial \beta_j \partial \beta_h} \right) = \sum_{i=1}^n x_{ij} x_{ih} \frac{\omega_i}{\phi g'(\mu_i)^2 V(\mu_i)}, \quad (5.4.9)$$

in quanto il secondo addendo delle derivate seconde di $\tilde{\ell}$ ha speranza matematica nulla.

Posto $W(\beta) = \text{diag} \left[\frac{\omega_i}{\phi g'(\mu_i)^2 V(\mu_i)} \right]$, si ha

$$\mathcal{J}(\beta) = X' W(\beta) X. \quad (5.4.10)$$

Osserviamo che poiché $W(\beta)$ è una matrice definita positiva e X ha rango pieno, $\mathcal{J}(\beta)$ è definita positiva (v. Appendice A). Inoltre, $\mathcal{J}(\beta)$ è la matrice delle varianze e covarianze del vettore aleatorio $\tilde{s}(\beta) = \left(\frac{\partial \tilde{\ell}}{\partial \beta_0}, \dots, \frac{\partial \tilde{\ell}}{\partial \beta_m} \right)'$. Infatti, dalla

$$\frac{\partial \tilde{\ell}(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} (Y_i - \mu_i) \frac{1}{g'(\mu_i) \text{var}(Y_i)} \quad (5.4.11)$$

e dall'indipendenza stocastica di Y_1, \dots, Y_n , si ha

$$\begin{aligned} \text{cov} \left(\frac{\partial \tilde{\ell}(\beta)}{\partial \beta_j}, \frac{\partial \tilde{\ell}(\beta)}{\partial \beta_h} \right) &= \text{cov} \left(\sum_{i=1}^n x_{ij} \frac{Y_i - \mu_i}{g'(\mu_i) \text{var}(Y_i)}, \sum_{k=1}^n x_{kh} \frac{Y_k - \mu_k}{g'(\mu_k) \text{var}(Y_k)} \right) \\ &= \sum_{i=1}^n \sum_{k=1}^n \frac{x_{ij} x_{kh}}{g'(\mu_i) \text{var}(Y_i) g'(\mu_k) \text{var}(Y_k)} \text{cov}(Y_i, Y_k) \\ &= \sum_{i=1}^n x_{ij} x_{ih} \frac{1}{g'(\mu_i)^2 \text{var}(Y_i)} = \sum_{i=1}^n x_{ij} x_{ih} \frac{\omega_i}{\phi g'(\mu_i)^2 V(\mu_i)}. \end{aligned}$$

Nel caso di collegamento canonico, dalla $g'(\mu_i) = 1/V(\mu_i)$, si verifica facilmente che

$$-H(\beta) = \mathcal{J}(\beta). \quad (5.4.12)$$

In tal caso, la matrice hessiana della log-verosimiglianza è definita negativa e quindi, come avevamo anticipato, la log-verosimiglianza è strettamente concava.

$$\textcircled{A} \quad \text{f}''(u_i) = -\frac{V''(u_i)}{V(u_i)^2} \Rightarrow \text{f}''(u_i)v(u_i) + f'(u_i)v'(u_i) = -\frac{V''(u_i)}{V(u_i)^2} v(u_i) + \frac{v'(u_i)}{V(u_i)} = 0$$

Fino a questo punto sono stati considerati gli elementi che servono per la stima del vettore dei parametri di regressione. Anche il parametro ϕ di dispersione, quando non è noto, può essere stimato con il metodo della massima verosimiglianza. Si tratta di risolvere l'equazione

$$\frac{\partial l(\hat{\boldsymbol{\theta}}, \phi; \mathbf{y})}{\partial \phi} = 0,$$

dove $\hat{\boldsymbol{\theta}}$ indica la stima del vettore dei parametri canonici, ottenuta come funzione della stima $\hat{\boldsymbol{\beta}}$ del vettore dei parametri di regressione. Nei GLM, però, ϕ è visto come un *parametro di disturbo*, mentre il *parametro di interesse*, oggetto delle analisi inferenziali, è il vettore dei parametri di regressione. Per tale motivo, spesso per ϕ sono utilizzati altri metodi di stima. Rimandiamo per tale aspetto al § 6.3, dove sono introdotti due stimatori per ϕ .

Le principali proprietà asintotiche dello stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ sono riportate nell'Appendice C. Qui sottolineiamo che, in forza di tali proprietà, da un punto di vista operativo, se il numero di osservazioni è "sufficientemente grande", si può supporre che lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ abbia distribuzione normale con matrice delle varianze e covarianze $[J(\hat{\boldsymbol{\beta}})]^{-1}$ o anche $[-H(\hat{\boldsymbol{\beta}})]^{-1}$, ottenute rispettivamente dalle matrici di informazione attesa e osservata calcolate nella stima $\hat{\boldsymbol{\beta}}$.

5.5 Metodi numerici per la stima dei parametri di regressione

Riprendiamo il sistema (5.4.4) delle equazioni di verosimiglianza

$$\sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{1}{g'(\mu_i) V(\mu_i)} = 0, \quad j = 0, \dots, m,$$

o, con scrittura vettoriale,

$$\mathbf{X}' \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}.$$

Come detto nel paragrafo precedente, per la risoluzione del sistema si ricorre a metodi numerici di tipo iterativo. Descriviamo i metodi di Newton-Raphson e *scoring* di Fisher.

Metodo di Newton-Raphson

Si considera un valore iniziale $\boldsymbol{\beta}^{(0)}$ per il vettore dei parametri di regressione. L'iterazione che fa passare dal valore $\boldsymbol{\beta}^{(k)}$ del passo k a $\boldsymbol{\beta}^{(k+1)}$ del passo successivo è

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} - [H(\boldsymbol{\beta}^{(k)})]^{-1} s(\boldsymbol{\beta}^{(k)}) \\ &= \boldsymbol{\beta}^{(k)} + [\mathbf{X}' \mathbf{W}_{oss}(\boldsymbol{\beta}^{(k)}) \mathbf{X}]^{-1} s(\boldsymbol{\beta}^{(k)}), \end{aligned} \tag{5.5.1}$$

dove $s(\boldsymbol{\beta}^{(k)})$ e $H(\boldsymbol{\beta}^{(k)}) = -\mathbf{X}' \mathbf{W}_{oss}(\boldsymbol{\beta}^{(k)}) \mathbf{X}$ (v. (5.4.7)) sono, rispettivamente, il vettore di punteggio e la matrice hessiana calcolati in $\boldsymbol{\beta}^{(k)}$. Per una giustificazione del metodo si veda l'Appendice D.

Metodo scoring di Fisher

È una variante del metodo di Newton-Raphson ottenuta sostituendo l'opposta della matrice hessiana $-H(\beta)$ con la matrice di informazione di Fisher $\mathcal{J}(\beta) = X'W(\beta)X$ (v. (5.4.10)). Quest'ultima è più facile da calcolare ed è sempre definita positiva.

Anche in questo caso si considera un valore iniziale $\beta^{(0)}$. L'iterazione dal passo k al passo successivo è ora

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + [\mathcal{J}(\beta^{(k)})]^{-1} s(\beta^{(k)}) \\ &= \beta^{(k)} + [X'W(\beta^{(k)})X]^{-1} s(\beta^{(k)}).\end{aligned}\quad (5.5.2)$$

Se la funzione di collegamento del modello è la canonica, per la (5.4.12), le (5.5.1), (5.5.2) coincidono.

Scriviamo in modo diverso la (5.5.2) per mostrare che nei GLM, il generico passo del procedimento *scoring* di Fisher equivale ad esprimere una stima dei minimi quadrati ponderati.

Dalla (5.4.5), la funzione di punteggio è $s(\beta) = X'DV^{-1}(y - \mu)$, dove ricordiamo che D , V e μ sono funzioni di β . Poniamo $z(\beta) = DV^{-1}(y - \mu)$, ovvero $z(\beta) = (z_1, \dots, z_n)'$ con

$$z_i = \frac{\omega_i}{\phi} (y_i - \mu_i(\beta)) \frac{1}{g'(\mu_i(\beta))V(\mu_i(\beta))}, \quad \in \mathbb{R}^m$$

dove $\mu_i(\beta) = g^{-1}(x'_i \beta)$. La funzione di punteggio può allora essere scritta come prodotto di X' e $z(\beta)$, si ha cioè

$$s(\beta) = X'z(\beta).$$

Poiché $\mathcal{J}(\beta) = X'W(\beta)X$, se poniamo $W^{(k)} = W(\beta^{(k)})$ e $z^{(k)} = z(\beta^{(k)})$, il secondo membro della (5.5.2) diventa

$$\begin{aligned}\beta^{(k)} + [X'W^{(k)}X]^{-1}X'z^{(k)} &= [X'W^{(k)}X]^{-1}[X'W^{(k)}X]\beta^{(k)} + [X'W^{(k)}X]^{-1}X'z^{(k)} \\ &= [X'W^{(k)}X]^{-1}X'W^{(k)}\left(X\beta^{(k)} + W^{(k)-1}z^{(k)}\right) \\ &= [X'W^{(k)}X]^{-1}X'W^{(k)}z^{*(k)},\end{aligned}$$

dove $z^{*(k)} = X\beta^{(k)} + W^{(k)-1}z^{(k)}$ è il vettore di componente i -esima

$$\begin{aligned}z_i^{*(k)} &= x'_i \beta^{(k)} + g'(\mu_i^{(k)})(y_i - \mu_i^{(k)}) \\ &= g(\mu_i^{(k)}) + g'(\mu_i^{(k)})(y_i - \mu_i^{(k)}),\end{aligned}$$

che rappresenta l'approssimazione del primo ordine di $g(y_i)$ con punto iniziale $\mu_i^{(k)} = \mu_i(\beta^{(k)})$.

L'iterazione nel procedimento *scoring* di Fisher può allora essere espressa dalla

$$\beta^{(k+1)} = [X'W^{(k)}X]^{-1}X'W^{(k)}z^{*(k)}.$$

Confrontando l'ultima espressione con la (4.3.3), si nota che il generico passo del procedimento equivale a calcolare la stima dei minimi quadrati ponderati utilizzando

come vettore dei dati $\mathbf{z}^{*(k)}$, che è detto *vettore degli pseudodati* o *vettore delle variabili dipendenti aggiustate*, e come matrice dei pesi $\mathbf{W}^{(k)}$. Pertanto, nel metodo *scoring* di Fisher, la stima dei parametri di regressione è ottenuta mediante un procedimento del tipo *minimi quadrati ponderati iterati*.

In convenienti ipotesi, le successioni che si ottengono dalle (5.5.1), (5.5.2) convergono alla soluzione del sistema delle equazioni di verosimiglianza. Naturalmente, da un punto di vista operativo, si deve considerare una condizione di arresto. Per esempio, si stabilisce di interrompere le iterazioni quando si abbia

$$\frac{\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|}{\|\boldsymbol{\beta}^{(k)}\|} \leq \varepsilon,$$

con $\varepsilon > 0$ prefissato e $\|\boldsymbol{\beta}\| = \sqrt{\sum_{j=0}^m \beta_i^2}$.

Per quanto riguarda la scelta del valore iniziale $\boldsymbol{\beta}^{(0)}$ del procedimento iterativo, qualora i valori osservati y_1, \dots, y_n appartengano al dominio della funzione di collegamento, si può per esempio fissare $\boldsymbol{\beta}^{(0)}$ pari alla stima dei minimi quadrati ordinari del modello di regressione lineare applicato al vettore dei dati $(g(y_1), \dots, g(y_n))$ e con matrice di regressione \mathbf{X}

$$\boldsymbol{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'g(\mathbf{y}).$$

Un'altra possibilità è fissare

$$\boldsymbol{\beta}^{(0)} = [\mathbf{X}'\mathbf{W}^{(0)}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}^{(0)}g(\mathbf{y}),$$

dove $\mathbf{W}^{(0)} = \text{diag}\left[\frac{\omega_i}{\phi g'(y_i)^2 V(y_i)}\right]$. Pertanto, $\boldsymbol{\beta}^{(0)}$ è la stima dei minimi quadrati ponderati ottenuta usando i valori osservati come stime iniziali delle speranze matematiche delle variabili risposta.

Osserviamo infine che le stime ottenute con i precedenti procedimenti numerici non dipendono dal parametro di dispersione. Infatti, nell'eseguire i prodotti $[\mathbf{H}(\boldsymbol{\beta}^{(k)})]^{-1}s(\boldsymbol{\beta}^{(k)})$ della (5.5.1) e $[\mathcal{J}(\boldsymbol{\beta}^{(k)})]^{-1}s(\boldsymbol{\beta}^{(k)})$ della (5.5.2) si elimina la dipendenza da ϕ .

Nota. Nell'ambito dei GLM, i problemi relativi all'esistenza e unicità della stima di massima verosimiglianza dei parametri di regressione, e al fatto che essa si possa determinare risolvendo il sistema delle equazioni di verosimiglianza, sono stati studiati da vari autori. Per particolari modelli sono state determinate le condizioni che garantiscono l'esistenza e l'unicità delle stime. Rimandiamo per tali aspetti alla letteratura specializzata (v. per esempio Fahrmeir, Tutz (2001), pag. 43). Spesso le condizioni non sono facili da verificare. Da un punto di vista pratico, conviene applicare il procedimento iterativo e verificare se si ha convergenza. Usualmente, il procedimento si arresta dopo poche iterazioni. Se il numero di iterazioni diventa elevato, ciò può essere dovuto ad una scelta inadeguata del valore iniziale $\boldsymbol{\beta}^{(0)}$ oppure al fatto che non esiste una stima di massima verosimiglianza interna allo spazio dei parametri. Quest'ultimo problema può essere superato se si dispone di molte osservazioni: i risultati asintotici sulle stime di

massima verosimiglianza garantiscono, infatti, che la probabilità che esista un vettore β , interno allo spazio parametrico, che rende massima la funzione di verosimiglianza tende a uno al divergere del numero di osservazioni (v. Appendice C). ♦

Nel prossimo paragrafo ci soffermiamo ad analizzare l'effetto dei diversi tipi di variabili esplicative sul previsore lineare e quindi sul vettore dei parametri di regressione.

5.6 Variabili esplicative e previsore lineare

Con riferimento alla classificazione delle variabili esplicative (v. § 3.2), analizziamo l'espressione degli addendi del previsore lineare corrispondenti ai diversi tipi di variabili. Nelle analisi che seguono, supponiamo che le colonne introdotte nella matrice X di regressione per le variabili di volta in volta considerate non siano combinazioni lineari delle rimanenti.

Variabili numeriche o quantitative

L'inserimento nel modello di una variabile numerica comporta l'introduzione di un unico parametro di regressione. Con riferimento alla variabile X_j e all' i -esima osservazione, se x_{ij} è la determinazione della variabile, l'addendo corrispondente nel previsore lineare η_i è $x_{ij}\beta_j$.

Se si considerano come variabili esplicative trasformate di variabili numeriche, ogni trasformata introdotta nel modello comporta l'inserimento di un parametro di regressione. Per esempio, se si introducono, oltre alla variabile X_j , anche le sue potenze di esponente due e tre, X_j^2 e X_j^3 , nel previsore lineare dell'osservazione i -esima si ha un addendo del tipo

$$x_{ij}\beta_1 + x_{ij}^2\beta_2 + x_{ij}^3\beta_3.$$

Osserviamo che in tal modo non si distrugge la linearità del previsore, perché essa è intesa rispetto ai parametri di regressione.

Si noti che, con l'inserimento di due variabili numeriche, X_1 , X_2 , non si tiene conto dell'influenza congiunta delle due variabili. Infatti, nel previsore lineare dell' i -esima osservazione, si ha un addendo del tipo

$$x_{i1}\beta_1 + x_{i2}\beta_2.$$

Pertanto, se la prima variabile esplicativa assume determinazione x_{i1} , il suo effetto è espresso dall'addendo $x_{i1}\beta_1$, indipendentemente dalla determinazione dell'altra variabile. Per tenere conto dell'effetto congiunto, si può introdurre nel modello, oltre alle variabili X_1 , X_2 , il prodotto X_1X_2 . Nel previsore lineare si ha allora un addendo del tipo

$$x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\beta_{12}.$$

Variabili di classificazione

Come indicato nel § 3.3, le variabili qualitative e le numeriche con determinazioni ripartite in livelli sono descritte da sequenze di variabili indicatori.

Sia C una variabile di classificazione con ℓ livelli. Poiché abbiamo supposto che il modello contenga l'intercetta, se C è codificata mediante ℓ variabili indicatori, allora le colonne della matrice di regressione X relative a tali variabili e all'intercetta sono linearmente dipendenti. Infatti, la prima colonna di X è somma delle colonne corrispondenti ai livelli della variabile di classificazione C . Pertanto, consideriamo la codificazione di C mediante $\ell-1$ variabili indicatori. L'inserimento della variabile comporta dunque l'introduzione di $\ell-1$ parametri di regressione. Indichiamo con $(X_1^{(C)}, \dots, X_{\ell-1}^{(C)})$ le variabili che codificano C e con $(x_{i1}^{(C)}, \dots, x_{i\ell-1}^{(C)})$, dove al più una componente è pari a uno e le rimanenti sono nulle, le determinazioni di tali variabili per l' i -esima osservazione. Nel previsore lineare si ha allora, in corrispondenza della variabile di classificazione C , un addendo del tipo

$$x_{i1}^{(C)}\beta_1^{(C)} + x_{i2}^{(C)}\beta_2^{(C)} + \dots + x_{i\ell-1}^{(C)}\beta_{\ell-1}^{(C)} = \begin{cases} \beta_j^{(C)} & \text{se } x_{ij}^{(C)} = 1 \\ 0 & \text{se } x_{ij}^{(C)} = 0, j = 1, \dots, \ell-1. \end{cases}$$

Pertanto, se le variabili esplicative del modello sono tutte di classificazione, il previsore lineare per l' i -esima osservazione è

- $\eta_i = \beta_0$, se l'osservazione appartiene alla *classe di riferimento* (tutte le variabili indicatori introdotte per la codificazione sono nulle),
- $\eta_i = \beta_0 + \beta_{j_1}^{(C_1)} + \dots + \beta_{j_s}^{(C_s)}$, se s variabili indicatori, corrispondenti a s variabili di classificazione, C_1, \dots, C_s , sono unitarie.

Esempio 5.6.1. Con riferimento ad un portafoglio di assicurati RCA, consideriamo le due variabili tariffarie di classificazione: *età* dell'assicurato con determinazioni ripartite in tre livelli 1, 2, 3; *potenza* fiscale del veicolo con determinazioni ripartite in due livelli 1, 2. Tali variabili siano codificate con le seguenti variabili indicatori

$$\begin{array}{ll} \text{età} & X_1^{(e)} = \begin{cases} 1 & \text{se età} = 1 \\ 0 & \text{altrimenti} \end{cases} \quad \text{e} \quad X_2^{(e)} = \begin{cases} 1 & \text{se età} = 2 \\ 0 & \text{altrimenti} \end{cases} \\ \text{potenza} & X_1^{(p)} = \begin{cases} 1 & \text{se potenza} = 1 \\ 0 & \text{altrimenti.} \end{cases} \end{array}$$

La classe di riferimento è dunque quella degli assicurati con *età* = 3, *potenza* = 2. L'espressione del previsore lineare per l' i -esima osservazione è

$$\eta_i = \beta_0 + x_{i1}^{(e)}\beta_1^{(e)} + x_{i2}^{(e)}\beta_2^{(e)} + x_{i1}^{(p)}\beta_1^{(p)}.$$

In dettaglio, il previsore lineare per le diverse classi tariffarie è riportato nella seguente tabella.

età	potenza	Previsore lineare
1	1	$\beta_0 + \beta_1^{(e)} + \beta_1^{(p)}$
1	2	$\beta_0 + \beta_1^{(e)}$
2	1	$\beta_0 + \beta_2^{(e)} + \beta_1^{(p)}$
2	2	$\beta_0 + \beta_2^{(e)}$
3	1	$\beta_0 + \beta_1^{(p)}$
3	2	β_0

Ci soffermiamo sul ruolo della classe di riferimento in relazione alle speranze matematiche delle variabili risposta, nel caso in cui le variabili tariffarie siano tutte di classificazione. Se, per esempio, la funzione di collegamento del GLM è l'identica, per tali speranze matematiche si ha

- $\mu_i = \beta_0$, se l'osservazione appartiene alla *classe di riferimento*,
- $\mu_i = \beta_0 + \beta_{j_1}^{(C_1)} + \dots + \beta_{j_s}^{(C_s)}$, se s variabili indicatrici sono unitarie.

Si ottiene quindi un modello tariffario additivo (v. § 2.3) in cui β_0 è il risarcimento atteso per rischio (o il numero atteso di sinistri o il risarcimento atteso per sinistro) per gli assicurati della classe tariffaria di riferimento e $\beta_{j_h}^{(C_h)}$ è la relatività additiva associata al livello j_h della variabile tariffaria C_h , che corrisponde alla variabile indicatrice $X_{j_h}^{(C_h)}$ e quindi alla condizione $X_{j_h}^{(C_h)} = 1$.

Se la funzione di collegamento del GLM è il logaritmo, si ha

- $\mu_i = e^{\beta_0}$, se l'osservazione appartiene alla *classe di riferimento*,
- $\mu_i = e^{\beta_0} e^{\beta_{j_1}^{(C_1)}} \dots e^{\beta_{j_s}^{(C_s)}}$, se s variabili indicatrici sono unitarie.

Si ottiene quindi un modello tariffario moltiplicativo (v. § 2.3) in cui $\exp(\beta_0)$ è il risarcimento atteso per rischio (o il numero atteso di sinistri o il risarcimento atteso per sinistro) per gli assicurati della classe tariffaria di riferimento e $\exp(\beta_{j_h}^{(C_h)})$ è la relatività moltiplicativa associata al livello j_h della variabile tariffaria C_h .

In entrambi i modelli, se $\beta_{j_h}^{(C_h)} > 0$, la modalità codificata dalla $X_{j_h}^{(C_h)} = 1$ rappresenta un fattore aggravante per la sinistrosità, rispetto alla sinistrosità dei rischi della classe tariffaria di riferimento; viceversa, se $\beta_{j_h}^{(C_h)} < 0$. Alla luce delle precedenti osservazioni, può essere opportuno scegliere come classe tariffaria di riferimento la più numerosa. Ciò rende possibile confrontare direttamente i risarcimenti attesi per rischio (o i numeri attesi di sinistri o i risarcimenti attesi per sinistro) delle diverse classi tariffarie con quelli della classe tariffaria maggiormente rappresentata.

Osserviamo che una variabile con determinazioni ripartite in livelli numerati può essere trattata sia come variabile di classificazione sia come variabile numerica. Se è trattata come variabile di classificazione si stima un diverso parametro per ogni livello, se è trattata come numerica si introduce un solo parametro.

Nel § 3.3 è stato descritto l'effetto di interazione tra due variabili di classificazione, A in a livelli, B in b livelli. Per quanto riguarda la numerosità dei parametri, sempre nell'ipotesi che il modello contenga l'intercetta, osserviamo ora che se nel modello sono inseriti solo i due effetti principali A e B , in relazione a tali effetti il modello contiene $(a-1)+(b-1)$ parametri. Se è inserita invece l'interazione $A * B$, il modello contiene $ab-1$ parametri. Ricordiamo che l'interazione può essere introdotta considerando gli effetti principali e i prodotti delle variabili indicatrici che codificano gli effetti principali. In tale caso il modello contiene $a-1$ parametri associati alla variabile A , $b-1$ parametri associati alla variabile B e ulteriori $(a-1) \times (b-1)$ parametri: complessivamente si hanno ancora $ab-1$ parametri.

Esempio 5.6.2. Con riferimento alle variabili dell'Esempio 5.6.1, in cui l'*età*, in tre livelli, è codificata da $(X_1^{(e)}, X_2^{(e)})$, e la *potenza fiscale*, in due livelli, è codificata da $X_1^{(p)}$, l'interazione *età * potenza* presenta sei livelli e può essere codificata con cinque variabili binarie del tipo

$$X_{ij}^{(ep)} = \begin{cases} 1 & \text{se età} = i, \text{potenza} = j \\ 0 & \text{altrimenti.} \end{cases}$$

Il modello con *intercetta, età, potenza* contiene $1+2+1=4$ parametri. Il modello con *intercetta, età * potenza* contiene $1+5=6$ parametri. Il modello con *intercetta, età, potenza, età * potenza* contiene $1+2+1+2=6$ parametri: alle variabili che codificano *età* e *potenza* basta aggiungere i prodotti: $X_1^{(e)}X_1^{(p)} = X_{11}^{(ep)}$, $X_2^{(e)}X_1^{(p)} = X_{21}^{(ep)}$ (cfr. Esempio 3.3.4). ◆

Naturalmente, nel caso di più variabili si possono considerare interazioni delle variabili a due a due, a tre a tre, e così via. Occorre però tenere presente che l'inserimento di interazioni può condurre ad un modello con un numero molto elevato di parametri.

Un modello che non contiene interazioni tra variabili è detto *modello con effetti principali*; altrimenti è detto *modello con effetti di interazione*.

In presenza di interazioni, nella letteratura è spesso suggerito di introdurre nel modello anche gli effetti principali, in modo da potere meglio valutare l'apporto delle singole variabili e degli effetti congiunti. Analogamente, in presenza di termini polinomiali è indicato di introdurre anche tutte le potenze di ordine inferiore rispetto alla massima inserita.

Effetti congiunti tra variabili numeriche e variabili di classificazione

Per tenere conto dell'effetto congiunto tra una variabile quantitativa, o una sua trasformata, e una variabile di classificazione si possono introdurre nel modello i prodotti tra la variabile quantitativa e le variabili indicatorie che codificano i livelli della variabile di classificazione. Indichiamo con X_j la variabile numerica e con C la variabile di classificazione, codificata da $(X_1^{(C)}, \dots, X_{\ell-1}^{(C)})$. Se per l' i -esima osservazione la determinazione di X_j è x_{ij} e il livello della variabile C è h , $h \leq \ell-1$, (cioè $x_{ih}^{(C)} = 1$, dove $x_{ih}^{(C)}$ è la determinazione della variabile indicatorie $X_h^{(C)}$), indicato con β_{jh} il parametro associato alla variabile $X_j X_h^{(C)}$, nel previsore lineare si ha un addendo del tipo $x_{ij}\beta_{jh}$. Tali addendi sono anche detti *termini misti*. L'introduzione nel modello di una variabile numerica, di una variabile di classificazione, codificata con $\ell-1$ variabili indicatorie, e dei corrispondenti termini misti comporta l'introduzione di $1 + 2(\ell-1)$ parametri di regressione.

Offset

In alcuni casi, si vuole includere nel modello una variabile esplicativa con effetto noto. Ciò può essere realizzato introducendo nel previsore lineare η un termine noto ξ ,

$$\eta = X\beta + \xi.$$

Il vettore ξ è detto anche *termine offset*.

Con un termine *offset* si può introdurre, in modo semplice, un vincolo per alcuni parametri di regressione imponendo che assumano prefissati valori. Infatti, considerata la ripartizione $\beta' = (\beta'_0, \beta'_1)$ del vettore dei parametri di regressione e la corrispondente

ripartizione $X = (X_0, X_1)$ della matrice di regressione, il vincolo $\beta_1 = b_1$ comporta che il previsore lineare sia del tipo

$$\eta = X_0 \beta_0 + X_1 b_1,$$

dove $\xi = X_1 b_1$ è un termine *offset*.

La possibilità di introdurre un termine *offset* può inoltre essere utile per tenere conto di misure di esposizione (v. § 7.2).

In chiusura del paragrafo, osserviamo che quando si introducono in un modello più variabili esplicative può accadere che le colonne della matrice di regressione non siano linearmente indipendenti, come invece è stato supposto nella definizione dei GLM nel § 5.3. Per esempio, come abbiamo osservato sopra, in un modello con intercetta e una variabile di classificazione con ℓ livelli, se la variabile di classificazione è codificata con ℓ variabili indicatori, il rango della matrice di regressione è minore del numero delle colonne. In tali casi, per rientrare nella definizione, occorrerebbe depurare la matrice di regressione dalle colonne che sono combinazioni lineari delle rimanenti, altrimenti si avrebbe indeterminatezza nella stima dei parametri. L'indeterminatezza può essere risolta, in modo equivalente, introducendo opportuni vincoli sui parametri, per esempio fissando pari a zero i parametri corrispondenti alle colonne della matrice di regressione combinazioni lineari delle altre. I pacchetti statistici che trattano i GLM effettuano automaticamente la precedente operazione. Per gli sviluppi teorici che seguono continuiamo a supporre che la matrice del modello abbia rango massimo.

5.7 Modelli lineari generalizzati in SAS

Per i GLM sono stati sviluppati diversi *software* statistici, che consentono di effettuare la stima dei parametri, la selezione delle variabili e analisi relative all'inferenza. Per le applicazioni alla tariffazione, che sono oggetto dei prossimi capitoli, facciamo riferimento al modulo disponibile in SAS. In questo paragrafo descriviamo brevemente la procedura *genmod* di STAT/SAS. Ci soffermiamo soltanto sugli aspetti principali; per una descrizione più dettagliata ed approfondita rimandiamo al Manuale del *SAS Institute* riportato in bibliografia.

I dati siano memorizzati in un *data set* di nome *archivio*. Esso contenga, per ogni osservazione, la determinazione della variabile risposta, il peso della distribuzione e le determinazioni di alcune variabili esplicative. Indichiamo con *risposta* il nome della variabile risposta, con *peso* il nome della variabile con valori i pesi, con X_1, \dots, X_n le variabili esplicative quantitative e con C_1, \dots, C_m le variabili esplicative di classificazione.

Illustriamo come si dichiara un GLM in SAS, nel caso in cui si attribuiscano alle variabili risposta distribuzioni di una famiglia esponenziale lineare predefinita e si consideri una funzione di collegamento anch'essa scelta tra le predefinite. Le distribuzioni e le funzioni di collegamento predefinite, per i GLM unidimensionali, insieme con le parole chiave che le individuano, sono riportate rispettivamente nella Tabella 5.4 e nella Tabella 5.5.

Il simbolo Φ che compare nella funzione di collegamento *probit* indica la funzione di ripartizione della normale *standard*. Le funzioni di collegamento *logit*, *probit* e *cloglog* (*complementary log-log*) sono appropriate per le distribuzioni bernoulliana e binomiale.

Tabella 5.4. Distribuzioni predefinite in SAS

Distribuzione	Parola chiave
Normale	normal
Poisson	poisson
Binomiale scalata	binomial
Binomiale negativa	negbin
Gamma	gamma
Gaussiana inversa	igaussian

Tabella 5.5. Funzioni di collegamento predefinite in SAS

Funzione di collegamento	Parola chiave
μ	identity
$\log\mu$	log
$\log(\mu/(1-\mu))$	logit
$\Phi^{-1}(\mu)$	probit
$\log\{-\log(1-\mu)\}$	cloglog
μ^γ , se $\gamma \neq 0$; $\log\mu$, se $\gamma = 0$	power(γ)

Veniamo alle istruzioni della procedura genmod.

```
proc genmod data = archivio;
  class C1...Cm;
  model risposta = X1...Xn C1...Cm / dist = parola chiave
    link = parola chiave
    <opzioni>;
  weight peso;
run;
```

Nell'istruzione **model** si precisa che *risposta* è la variabile risposta e X_1, \dots, X_n , C_1, \dots, C_m sono le variabili esplicative del modello.

Nell'istruzione **class** devono essere elencate le variabili di classificazione (*classification variables*). La procedura codifica ciascuna di tali variabili mediante le variabili indicatorie che ne identificano i livelli e inserisce nella matrice X di regressione le colonne corrispondenti. Una variabile con ℓ livelli è codificata con ℓ variabili indicatorie. Se una variabile esplicativa presente nell'istruzione **model** non compare nell'istruzione **class**, è trattata come variabile quantitativa ed in corrispondenza ad essa compare dunque un'unica colonna nella matrice X .

La procedura introduce in modo automatico un'intercetta, ovvero la prima colonna della matrice X è di termini tutti unitari. La matrice di regressione costruita dalla procedura **genmod** non ha generalmente rango pieno. Se nella matrice X compare una colonna combinazione lineare di colonne precedenti, il corrispondente parametro di regressione non è stimato ed è posto pari a zero. Poiché per una variabile di classificazione con ℓ livelli vengono inserite nella matrice di regressione ℓ colonne, corrispondenti alle variabili indicatorie che ne rappresentano i diversi livelli, allora la colonna che rappresenta

l'ultimo livello è combinazione lineare delle precedenti $\ell - 1$ e della colonna che rappresenta l'intercetta. Pertanto, il corrispondente parametro di regressione è posto uguale a zero. Ciò equivale a codificare una variabile di classificazione con determinazioni in ℓ livelli mediante $\ell - 1$ variabili indicatrici; inoltre, in modo automatico, si sceglie come livello di riferimento quello di valore massimo, nell'ordine naturale se i livelli sono numerati, nell'ordine lessicografico se i livelli sono denominati con stringhe di caratteri.

Tra le opzioni che compaiono dopo la barra (/) presentiamo, per il momento, le tre seguenti: **dist** che specifica la distribuzione di probabilità assegnata alle variabili risposta, scelta tra le predefinite; **link** che specifica la funzione di collegamento, tra le predefinite; **noint** che richiede la stima di un modello senza l'intercetta. Se l'opzione **link** non compare, è usata la funzione canonica di collegamento. Altre opzioni saranno illustrate in seguito. In particolare, un'opzione consente di scegliere lo stimatore per il parametro di dispersione. Se tale opzione non è indicata, la stima del parametro di dispersione è effettuata con il metodo della massima verosimiglianza. A questo proposito, segnaliamo che **genmod** non fornisce direttamente la stima di ϕ , bensì di un parametro di scala (**scale parameter**) che è, in generale, trasformato di ϕ . La relazione tra il parametro di scala e ϕ per i diversi tipi di distribuzione è riportato nella Tabella 5.6. Per le distribuzioni di Poisson, binomiali scalate e binomiali negative il parametro di dispersione è noto, si ha $\phi = 1$, e il parametro di scala è posto uguale a 1. Per le distribuzioni binomiali negative, se il parametro α (v. Tabella 5.1) non è dato, la procedura stima con il metodo della massima verosimiglianza α^{-1} , tale parametro è denominato **dispersion parameter**.

Tabella 5.6. Parametri di scala e di dispersione

Distribuzione	Parametro di scala
Normale	$\sqrt{\phi}$
Gamma	$1/\phi$
Gaussiana inversa	$\sqrt{\phi}$

L'istruzione **weight** identifica la variabile del *data set* i cui valori devono essere utilizzati come pesi delle distribuzioni attribuite alle variabili risposta. Se i pesi sono tutti unitari non occorre dichiararli.

L'**output** della procedura fornisce: informazioni riassuntive sul modello; alcuni elementi che consentono di valutare la bontà di adattamento ai valori osservati; la stima dei parametri di regressione e del parametro di scala. Per ottenere tra i risultati anche i valori attesi stimati, basta inserire l'opzione **predicted** nell'istruzione **model**.

Esempio 5.7.1. Modello per i numeri annui di sinistri

I dati, utilizzati unicamente a fini illustrativi in questo e negli esempi che seguono, sono relativi ad un portafoglio di 172'161 polizze RCA osservate per un anno. Come già segnalato nel § 2.1, supponiamo che il risarcimento dei danni sia effettuato con la procedura ordinaria, secondo la quale il danneggiato è risarcito dalla compagnia che copre il responsabile del sinistro. In tale caso, per ciascun assicurato, le variabili risposta rilevanti per la tariffazione sono il risarcimento totale dovuto ai terzi danneggiati, secondo

le modalità contrattuali, il numero di sinistri causati dall'assicurato nel periodo di copertura, i risarcimenti per sinistro. Nel § 10.7 sono indicate le variabili tramite le quali si può descrivere la prestazione dell'assicuratore nello schema di risarcimento diretto attualmente in vigore in Italia.

Il portafoglio è ripartito in 12 classi individuate dalle determinazioni di due variabili tariffarie di classificazione: età dell'assicurato e potenza del veicolo in cavalli fiscali (CV), con le seguenti ripartizioni delle determinazioni in livelli

età	18-22, 23-26, 27-43, maggiore di 43 (anni),
potenza	8-12, 13-17, maggiore di 17 (CV).

I dati sono memorizzati in un *data set* di SAS denominato *polizze* che contiene per ogni polizza (tra parentesi, in corsivo, sono indicati i nomi delle corrispondenti variabili): l'età dell'assicurato (*eta*), la potenza fiscale del veicolo (*potf*), l'esposizione o rischio/anno (*espo*), il numero di sinistri causati nel periodo di osservazione (*nsin*), l'importo totale di danno nel periodo di osservazione (*dannotot*). Sono inoltre noti gli importi di danno per sinistro, memorizzati in un altro *data set* di nome *danni*.

Il numero totale di sinistri per le polizze del portafoglio è 12.691, l'esposizione totale è 123.282, il danno totale è 38.073.000 euro.

Per l'analisi del numero annuo di sinistri che colpiscono un rischio, consideriamo qui solamente le polizze con esposizione unitaria, le modalità per trattare dati con esposizioni diverse sono illustrate nel Capitolo 7. Sia *polizze1* il nome del *data set* che contiene le 77.579 osservazioni relative a tali polizze che hanno causato complessivamente 6.885 sinistri. La tabella seguente fornisce un riassunto dei numeri annui di sinistri in *polizze1*.

Media (Frequenza sinistri)	Scarto quadratico medio	Valore minimo	Valore massimo
0,088748	0,310389	0	5

Al fine di stimare una distribuzione di Poisson per il numero di sinistri che colpiscono un rischio in un anno, distribuzione assunta comune per i rischi di una medesima classe tariffaria, definiamo un GLM per i numeri aleatori

$$M_i \text{ numero annuo di sinistri per l'}i\text{-esimo assicurato}.$$

La struttura del modello è la seguente

- *variabili risposta*: M_i stocasticamente indipendenti con distribuzioni di Poisson, $E(M_i) = \text{var}(M_i) = \mu_i$,
- *variabili esplicative*: *eta*, *potf*,
- *funzione di collegamento*: $g = \log$.

Con il seguente programma si ottiene la stima del modello in SAS.

```
proc genmod data = polizze1;
  class eta potf;
  model nsin = eta potf / dist = poisson
                        link = log;
run;
```

È stata scelta la funzione di collegamento logaritmo che è la canonica per la distribuzione di Poisson e che conduce ad un modello moltiplicativo. Poiché il logaritmo è il collegamento canonico, l'opzione **link** può essere omessa.

Il numero di osservazioni utilizzate è 77.579; il rango della matrice di regressione è 6.

La Tabella 5.7 riporta l'*output* della procedura che è ripartito in quattro sezioni. Nella prima, denominata **Model Information**, sono indicati: il nome del *data set* analizzato dalla procedura (**Data Set**), la famiglia di distribuzioni assegnate alle variabili risposta (**Distribution**), la funzione di collegamento del modello (**Link Function**), il nome della variabile risposta (**Dependent Variable**), il nome della variabile che ha come valori i pesi (**Scale Weight Variable**), in questo caso assente in quanto i pesi sono tutti unitari, e il numero di osservazioni utilizzate (**Observations Used**). Nella seconda sezione, denominata **Class Level Information**, sono riportate informazioni sulle variabili di classificazione inserite nel modello con i rispettivi nomi, numeri di livelli (**Levels**) e valori (**Values**). Nella sezione **Criteria For Assessing Goodness Of Fit** sono riportati i valori di statistiche che consentono di valutare la bontà di adattamento. Tali statistiche sono descritte nel § 6.2. Con riferimento alla sezione **Analysis Of Parameter Estimates**, per il momento ci soffermiamo sulle stime dei parametri; per i rimanenti elementi, che forniscono indicazioni sulla significatività dei parametri e sull'affidabilità delle stime, rimandiamo ai paragrafi 6.4 e 6.6. I nomi assegnati ai parametri sono evidenziati nella colonna **Parameter**, le stime nella colonna **Estimate**.

Si nota che, come già detto, la procedura fissa come livelli di riferimento delle due variabili esplicative i livelli massimi, *eta* > 43, *potf* > 17; i corrispondenti valori dei parametri di regressione sono posti pari a zero. Il parametro di scala non è stimato ed è posto uguale a 1.

La stima del numero atteso annuo di sinistri per un assicurato della classe tariffaria di riferimento è, con arrotondamento alla quarta cifra decimale per le stime dei parametri, $\exp(-2,2802) = 0,1023$. Rispetto agli assicurati con età al livello di riferimento, gli assicurati con età inferiore ai 27 anni, presentano sovra-sinistrosità. In particolare, la relatività per gli assicurati con età minore o uguale a 22 anni è $\exp(0,4489) = 1,5666$, quella per gli assicurati con età tra 23 e 26 anni è $\exp(0,2101) = 1,2338$. Gli assicurati con età tra 27 e 43 anni, presentano invece sotto-sinistrosità rispetto a quelli con età al livello di riferimento: la relatività corrispondente è $\exp(-0,1385) = 0,8707$.

Per quanto riguarda l'altra variabile esplicativa, gli assicurati con autoveicoli di potenza fiscale non superiore a 17 CV, presentano sotto-sinistrosità rispetto a quelli del livello di riferimento. La relatività per gli autoveicoli con potenza minore di 13 CV è $\exp(-0,2479) = 0,7804$, per gli autoveicoli con potenza tra 13 CV e 17 CV è $\exp(-0,1157) = 0,8907$.

Nella Tabella 5.8 sono riassunti, per le diverse classi tariffarie, i dati relativi ai numeri di polizze e di sinistri, le frequenze sinistri osservate e i numeri attesi annui di sinistri stimati dal modello.

Tabella 5.7. Risultati del modello per i numeri di sinistri

The GENMOD Procedure												
Model Information												
Data Set						WORK.POLIZZEL						
Distribution						Poisson						
Link Function						Log						
Dependent Variable						nsin						
Observations Used						77579						
Class Level Information												
		Class	Levels	Values								
		eta	4	18-22	23-26	27-43	>43					
		potf	3	8-12	13-17	>17						
Criteria For Assessing Goodness Of Fit												
Criterion		DF	Value		Value/DF							
Deviance		78E3	34690.5238		0.4472							
Scaled Deviance		78E3	34690.5238		0.4472							
Pearson Chi-Square		78E3	83923.9230		1.0819							
Scaled Pearson X2		78E3	83923.9230		1.0819							
Log Likelihood			-23461.5448									
Algorithm converged.												
Analysis Of Parameter Estimates												
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq						
Intercept	1	-2.2802	0.0293	-2.3377 -2.2227	6041.05	<.0001						
eta 18-22	1	0.4489	0.0505	0.3498 0.5479	78.93	<.0001						
eta 23-26	1	0.2101	0.0409	0.1299 0.2902	26.39	<.0001						
eta 27-43	1	-0.1385	0.0267	-0.1908 -0.0862	26.92	<.0001						
eta >43	0	0.0000	0.0000	0.0000 0.0000	.	.						
potf 8-12	1	-0.2479	0.0350	-0.3165 -0.1793	50.14	<.0001						
potf 13-17	1	-0.1157	0.0314	-0.1773 -0.0541	13.56	0.0002						
potf >17	0	0.0000	0.0000	0.0000 0.0000	.	.						
Scale	0	1.0000	0.0000	1.0000 1.0000	.	.						

NOTE: The scale parameter was held fixed.

Tabella 5.8. Modello per i numeri di sinistri: frequenze osservate e valori attesi stimati

età	potf	Numero di polizze	Numero di sinistri	Frequenza sinistri osservata	Numero atteso stimato
18-22	8-12	1235	150	0,121457	0,125028
18-22	13-17	1854	276	0,148867	0,142692
18-22	>17	200	25	0,125000	0,160196
23-26	8-12	2064	191	0,092539	0,098468
23-26	13-17	3721	423	0,113679	0,112380
23-26	>17	940	126	0,134043	0,126165
27-43	8-12	9240	626	0,067749	0,069491
27-43	13-17	16248	1315	0,080933	0,079309
27-43	>17	6787	594	0,087520	0,089037
>43	8-12	11568	956	0,082642	0,079811
>43	13-17	17013	1507	0,088579	0,091088
>43	>17	6709	696	0,103741	0,102261

I valori stimati sono generalmente molto vicini ai valori osservati, tranne che in alcune classi con un numero piuttosto basso di osservazioni. In particolare, lo scarto relativo

$$\frac{|frequenza\ osservata - numero\ atteso\ stimato|}{frequenza\ osservata}$$

è piuttosto elevato, superiore al 28%, nella classe con età 18-22 anni e potenza > 17 CV, in cui sono stati riportati solo 25 sinistri. Scarti relativi elevati, attorno al 6%, si hanno inoltre nelle due classi con età 23-26 e potenza 8-12 o > 17. Si noti l'effetto di perequazione realizzato dal modello. ♦

Esempio 5.7.2. Modello per i danni per sinistro

Il *data set* di nome *danni* contiene le osservazioni relative agli importi di danno corrispondenti ai 12'691 sinistri causati dai 172'161 rischi del portafoglio introdotto nell'Esempio 5.7.1. I dati disponibili per ogni osservazione sono gli importi di danno per sinistro (*danno*) e le determinazioni delle due variabili tariffarie, età dell'assicurato (*eta*) e potenza fiscale del veicolo (*potf*). I valori delle statistiche riportati nella seguente tabella forniscono una sintesi degli importi osservati.

Media (Danno medio per sinistro)	Scarto quadratico medio	Coefficiente di asimmetria	Percentile 0,25	Mediana	Percentile 0,75	Percentile 0,99
3000	10310,72	22,08	718,11	1617,55	2545,72	24236,36

Si vuole stimare una distribuzione gamma per il danno per sinistro. Si assume che la distribuzione sia la medesima per i rischi di una stessa classe tariffaria. Definiamo dunque un GLM per i numeri aleatori

Y_i importo del danno provocato dall'*i*-esimo sinistro.

La struttura del modello è la seguente

- *variabili risposta*: Y_i stocasticamente indipendenti con distribuzioni gamma, $E(Y_i) = \mu_i$, $var(Y_i) = \phi V(\mu_i) = \phi \mu_i^2$,
- *variabili esplicative*: *eta*, *potf*,
- *funzione di collegamento* $g = \log$.

In SAS si ha la seguente specificazione del modello.

```
proc genmod data = danni;
  class eta potf;
  model danno = eta potf / dist = gamma
    link = log;
run;
```

È stata scelta la funzione di collegamento logaritmo e non la canonica, che per la distribuzione gamma è $g(\mu) = -1/\mu$, in quanto quest'ultima richiederebbe di porre vincoli sui parametri di regressione per garantire che il previsore lineare sia negativo e quindi che la speranza matematica sia positiva. Si preferisce invece fissare una funzione di collegamento che conduca a valori positivi della speranza matematica, senza imporre vincoli ai parametri. La scelta del logaritmo è inoltre dettata dal fatto che spesso, in pratica, si adottano modelli tariffari di tipo moltiplicativo.

Tabella 5.9. Risultati del modello per i danni per sinistro

The GENMOD Procedure										
Model Information										
Data Set						WORK.DANNI				
Distribution						Gamma				
Link Function						Log				
Dependent Variable						danno				
Observations Used						12691				
Class Level Information										
Class	Levels	Values								
eta	4	18-22 23-26 27-43 >43								
potf	3	8-12 13-17 >17								
Criteria For Assessing Goodness Of Fit										
Criterion	DF	Value								
Deviance	13E3	18019.4477								
Scaled Deviance	13E3	14987.6885								
Pearson Chi-Square	13E3	130784.0698								
Scaled Pearson X2	13E3	108779.7439								
Log Likelihood		-113985.0415								
Algorithm converged.										
Analysis Of Parameter Estimates										
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq				
Intercept	1	8.0565	0.0228	8.0117 8.1012	124485	<.0001				
eta 18-22	1	0.3231	0.0367	0.2512 0.3950	77.52	<.0001				
eta 23-26	1	0.2228	0.0320	0.1600 0.2856	48.40	<.0001				
eta 27-43	1	0.0933	0.0219	0.0503 0.1363	18.10	<.0001				
eta >43	0	0.0000	0.0000	0.0000 0.0000	.	.				
potf 8-12	1	-0.3504	0.0283	-0.4058 -0.2950	153.57	<.0001				
potf 13-17	1	-0.1313	0.0241	-0.1784 -0.0841	29.78	<.0001				
potf >17	0	0.0000	0.0000	0.0000 0.0000	.	.				
Scale	1	0.8318	0.0090	0.8142 0.8497						

NOTE: The scale parameter was estimated by maximum likelihood.

Tabella 5.10. Modello per i danni per sinistro: dati medi e valori attesi stimati

eta	potf	Numero di sinistri	Danno totale	Danno medio osservato	Danno atteso stimato
18-22	8-12	261	750004	2873,58	3069,10
18-22	13-17	685	2766064	4038,05	3820,87
18-22	>17	135	490993,9	3636,99	4356,94
23-26	8-12	330	773416,6	2343,69	2776,26
23-26	13-17	873	2602503	2981,10	3456,30
23-26	>17	299	1854123	6201,08	3941,22
27-43	8-12	989	2375925	2402,35	2439,07
27-43	13-17	2407	7527951	3127,52	3036,52
27-43	>17	1243	4105714	3303,07	3462,54
>43	8-12	1408	3312557	2352,67	2221,80
>43	13-17	2654	7365764	2775,34	2766,02
>43	>17	1407	4147986	2948,11	3154,09

Con la precedente dichiarazione, il parametro ϕ di dispersione è stimato con il metodo della massima verosimiglianza. Nell'output non è riportata la stima di ϕ , ma la stima del parametro di scala che per la gamma è ϕ^{-1} .

Il numero di osservazioni è 12.691 e il rango della matrice di regressione è 6. Nella Tabella 5.9 è riportato l'output della procedura.

I dati relativi ai numeri di sinistri e ai danni totali, i danni medi per sinistro osservati e i danni attesi stimati, per le diverse classi tariffarie, sono riassunti nella Tabella 5.10.

La stima del danno atteso per sinistro per un assicurato della classe tariffaria di riferimento è $\exp(8,0565) = 3154,23$ euro (la differenza con il danno atteso della tabella è dovuta alle approssimazioni). Rispetto agli assicurati con età al livello di riferimento, tutti gli assicurati presentano sovraincidenza, con relatività decrescenti al crescere del livello di età.

Per quanto riguarda l'altra variabile esplicativa, gli assicurati con autoveicoli di potenza fiscale non superiore a 17 CV presentano sotto-sinistrosità rispetto a quelli del livello di riferimento.

Come nell'esempio sui numeri di sinistri, si notano un buon accostamento dei valori stimati ai valori osservati e l'effetto di perequazione realizzato dal modello. Anche qui gli scarti relativi più elevati si hanno in corrispondenza delle classi con relativamente poche osservazioni. Lo scarto maggiore, superiore al 36%, si ha nella classe con età 23-26 e potenza > 17, in cui il danno medio è molto elevato rispetto ai valori osservati nelle altre classi.

Per il parametro di dispersione si ottiene $\hat{\phi} = 1,2022$. Si noti che la stima del parametro di scala, che è il parametro di forma della distribuzione gamma, è minore di uno. Pertanto, le distribuzioni stimate per le diverse classi tariffarie sono di tipo gamma con funzione di densità superiormente illimitata e decrescente. Ciò è conseguenza della notevole dispersione dei dati (si veda la tabella con i valori delle statistiche di sintesi). Rimane tuttavia aperto il problema dell'analisi dell'adeguatezza del modello per i dati in esame. Le tecniche a tale fine sono illustrate nel Capitolo 6 (v. anche § 8.3). ◆

Esempio 5.7.3. Con riferimento all'Esempio 5.7.2, se una delle variabili esplicative, per esempio l'età dell'assicurato, è trattata come numerica anziché come variabile di classificazione, ferme restando le altre ipotesi del modello, per la stima in SAS si ha la seguente specificazione.

```
proc genmod data = danni;
  class potf;
  model danno = eta potf / dist = gamma
    link = log;
run;
```

In corrispondenza alla variabile *eta* c'è ora un unico parametro. Le stime, approssimate alla quarta cifra decimale, sono riportate nella tabella che segue.

Parameter	Estimate
Intercept	8.3547
eta	-0.0053
potf 8-12	-0.3338
potf 13-17	-0.1178
potf > 17	0.0000
Scale	0.8297

Dunque, per esempio, il valore stimato del danno atteso per sinistro per un assicurato di 23 anni con autoveicolo di potenza fiscale nella classe 13-17 CV è, approssimativamente, $\exp(8,3547 + 23 \times (-0,0053) - 0,1178) = 3344,26$.

Con la seguente specificazione si tiene invece anche conto dell'effetto congiunto delle due variabili.

```
proc genmod data = danni;
  class potf;
  model danno = eta potf eta*potf/dist = gamma
    link = log;
run;
```

In questo caso, oltre alle stime dei parametri relativi agli effetti principali, si hanno le stime dei termini misti.

Parameter	Estimate
Intercept	8.6981
eta	-0.0136
potf 8-12	-0.9442
potf 13-17	-0.4679
potf > 17	0.0000
eta* potf 8-12	0.0144
eta* potf 13-17	0.0085
eta* potf >17	0.0000
Scale	0.8328

Il valore stimato del danno atteso per sinistro per un assicurato di 23 anni con autoveicolo di potenza fiscale nella classe 13-17 CV è ora, approssimativamente, $\exp(8,6981 + 23 \times (-0,0136) - 0,4679 + 23 \times 0,0085) = 3337,24$. ♦

La procedura genmod consente di utilizzare funzioni di collegamento definite dall'utente tramite le istruzioni fwdlink e invlink nelle quali si indicano le espressioni della funzione di collegamento e della sua inversa, rispettivamente. Le istruzioni sono riportate qui di seguito.

```
proc genmod data = archivio;
  class C1...Cm;
  fwdlink link = espressione di g in funzione di _MEAN_;
  invlink ilink = espressione di g-1 in funzione di _XBETA_;
  model risposta = X1...Xn C1...Cm / dist = parola chiave
    <opzioni>;
  weight peso;
run;
```

Nel precedente programma link e ilink sono nomi di variabili che identificano la funzione di collegamento e la sua inversa, i simboli _MEAN_, _XBETA_ indicano invece variabili automaticamente definite che rappresentano la speranza matematica delle variabili risposta ed il previsore lineare. Per esempio, la dichiarazione

```
fwdlink link = log(_MEAN_);
invlink ilink = exp(_XBETA_);
```

equivale a dichiarare link = log come opzione nell'istruzione model.

È possibile anche specificare un modello con distribuzione appartenente ad una famiglia esponenziale lineare non predefinita. In tal caso si devono indicare la funzione di

varianza e la devianza della distribuzione, che è definita in seguito nel § 6.2. Per un esempio si veda, al § 8.4, la dichiarazione di un modello con distribuzione con funzione di varianza di tipo potenza.

5.8 Modelli con dati individuali e con dati raggruppati

Fino a questo punto è stato implicitamente supposto che i dati fossero *individuali* ovvero che, per ogni i , la determinazione della variabile risposta y_i e il vettore delle determinazioni delle variabili esplicative x_i corrispondessero ad un'unica unità statistica. Può accadere, ed anzi questa è la situazione normale nella tariffazione, che lo stesso vettore di determinazioni delle variabili esplicative sia comune a più unità statistiche. In tale caso i dati possono essere *raggruppati* ottenendo un GLM che, ai fini della stima dei parametri di regressione, è equivalente ad uno con dati individuali. In breve, nel modello con dati individuali la matrice di regressione contiene le determinazioni delle variabili esplicative per ogni singola unità statistica, nel modello con dati raggruppati contiene solo righe di determinazioni delle variabili esplicative diverse tra loro. Si tratta poi, nel secondo caso, di definire in modo opportuno le variabili risposta. Alla descrizione dettagliata dei due modelli, premettiamo il seguente teorema.

Teorema 5.8.1. I numeri aleatori Y_1, \dots, Y_n siano stocasticamente indipendenti con distribuzioni appartenenti ad una medesima famiglia esponenziale lineare. La funzione di probabilità o di densità di Y_i , $i = 1, \dots, n$, sia

$$f(y; \theta, \phi, \omega_i) = \exp\left\{\frac{\omega_i}{\phi}[y\theta - b(\theta)]\right\} c(y, \phi, \omega_i).$$

Posto $\omega = \sum_{i=1}^n \omega_i$, il numero aleatorio

$$Y = \sum_{i=1}^n \frac{\omega_i}{\omega} Y_i$$

ha distribuzione appartenente alla stessa famiglia esponenziale lineare delle distribuzioni di Y_1, \dots, Y_n , con

$$f(y; \theta, \phi, \omega) = \exp\left\{\frac{\omega}{\phi}[y\theta - b(\theta)]\right\} c(y, \phi, \omega).$$

Nota. Prima di passare alla dimostrazione, osserviamo che le distribuzioni dei numeri aleatori Y_1, \dots, Y_n hanno la stessa funzione cumulante b e gli stessi parametri θ e ϕ . Al variare di i , possono invece variare i pesi. La distribuzione del numero aleatorio Y , media ponderata di Y_1, \dots, Y_n con pesi $\omega_1, \dots, \omega_n$, ha ancora b come funzione cumulante e θ e ϕ come parametri canonico e di dispersione. Il peso della distribuzione è la somma dei pesi. ♦

DIMOSTRAZIONE. Consideriamo la funzione generatrice dei momenti di Y . Posto $k_i = \omega_i / \omega$, per l'indipendenza stocastica di Y_1, \dots, Y_n , si ha

$$m_Y(t) = E[e^{tY}] = E\left[\exp\left(t \sum_{i=1}^n k_i Y_i\right)\right] = \prod_{i=1}^n E[\exp(t k_i Y_i)] = \prod_{i=1}^n m_{Y_i}(t k_i).$$

Ricordando l'espressione (5.2.6) della funzione generatrice dei momenti delle distribuzioni di una famiglia esponenziale lineare, si ottiene

$$\left| \begin{aligned} m_Y(t) &= \prod_{i=1}^n \exp \left\{ \frac{\omega_i}{\phi} [b(\theta + t k_i \phi / \omega_i) - b(\theta)] \right\} = \prod_{i=1}^n \exp \left\{ \frac{\omega_i}{\phi} [b(\theta + t \phi / \omega) - b(\theta)] \right\} \\ &= \exp \left\{ \sum_{i=1}^n \frac{\omega_i}{\phi} [b(\theta + t \phi / \omega) - b(\theta)] \right\} = \exp \left\{ \frac{\omega}{\phi} [b(\theta + t \phi / \omega) - b(\theta)] \right\}. \end{aligned} \right|$$

L'espressione a ultimo membro è la funzione generatrice dei momenti di una distribuzione della famiglia esponenziale lineare con funzione cumulante b , parametro canonico θ , parametro di dispersione ϕ e peso ω . Poiché la funzione generatrice dei momenti individua la distribuzione di probabilità, si ha la tesi. ♦

Modello con dati individuali

Per ogni osservazione, consideriamo il vettore delle variabili esplicative e la determinazione della variabile risposta. Se ci sono osservazioni cui corrispondono uguali determinazioni delle variabili esplicative, si può ordinare la matrice di regressione in modo da evidenziare i blocchi di righe uguali. Un blocco individua una classe di osservazioni con le stesse caratteristiche: nel caso della tariffazione, una classe tariffaria. Indichiamo con K il numero di blocchi in cui è ripartita la matrice di regressione X e con n_1, \dots, n_K i numeri di righe in ciascuno dei blocchi.

Con riferimento all' i -esima osservazione nella classe k , $k = 1, \dots, K$, $i = 1, \dots, n_k$, siano Y_{ki} la variabile risposta, y_{ki} il valore osservato di Y_{ki} e x_{kij} la determinazione della j -esima variabile esplicativa, $j = 0, \dots, m$. Siano inoltre \mathbf{Y} il vettore delle variabili risposta e \mathbf{y} il vettore dei valori osservati. In modo schematico, si ha

$$X = \left[\begin{array}{cccc} 1 & x_{111} & \cdots & x_{11m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n_11} & \cdots & x_{1n_1m} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k11} & \cdots & x_{k1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{kn_k1} & \cdots & x_{kn_km} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{K11} & \cdots & x_{K1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{Kn_K1} & \cdots & x_{Kn_Km} \end{array} \right] \left\{ \begin{array}{l} n_1 \\ \vdots \\ n_k \\ \vdots \\ n_K \end{array} \right\} \quad \mathbf{y} = \left[\begin{array}{c} y_{11} \\ \vdots \\ y_{1n_1} \\ \hline \vdots \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \\ \hline \vdots \\ \vdots \\ y_{K1} \\ \vdots \\ y_{Kn_K} \end{array} \right] \quad \mathbf{Y} = \left[\begin{array}{c} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \hline \vdots \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \\ \hline \vdots \\ \vdots \\ Y_{K1} \\ \vdots \\ Y_{Kn_K} \end{array} \right]$$

dove, fissati k, j , risulta $x_{kij} = x_{khj}$, per $1 \leq h \leq n_k$. Poniamo dunque $x_{kj} = x_{kij}$, per $i = 1, \dots, n_k$, e $\mathbf{x}'_k = (x_{k0}, \dots, x_{km})$.

Definiamo un GLM per il precedente problema. I numeri aleatori Y_{ki} , $k = 1, \dots, K$, $i = 1, \dots, n_k$, siano stocasticamente indipendenti con distribuzioni appartenenti ad una medesima famiglia esponenziale lineare. Sia ω_{ki} il peso della distribuzione di Y_{ki} . Siano inoltre $\eta_k = \mathbf{x}'_k \boldsymbol{\beta}$ il previsore lineare comune alle osservazioni della classe k ,

il stesso previsore delle medesime classi

$k = 1, \dots, K$, e g la funzione di collegamento. Poiché si ha $E(Y_{ki}) = g^{-1}(\mathbf{x}'_k \boldsymbol{\beta}) = \mu_k$, la speranza matematica di Y_{ki} non dipende da i , allora la distribuzione di Y_{ki} è del tipo

$$Y_{ki} \sim \exp\left\{ \frac{\omega_{ki}}{\phi} [y\theta_{ki} - b(\theta_{ki})] \right\} c(y, \phi, \omega_{ki}) = \exp\left\{ \frac{\omega_{ki}}{\phi} [y\theta_k - b(\theta_k)] \right\} c(y, \phi, \omega_{ki}),$$

dove il parametro canonico non dipende da i poiché riesce $\theta_{ki} = b'^{-1}(\mu_k) = \theta_k$. Pertanto, esso varia solo al variare della classe; il peso può invece variare con l'osservazione. Naturalmente, la funzione cumulante e il parametro di dispersione sono gli stessi per ogni osservazione.

Il modello così ottenuto è detto *modello con dati individuali*.

► Modello con dati raggruppati

Con riferimento al modello sopra descritto, indichiamo con X^{cum} la matrice $K \times p$ ottenuta dalla X considerando una sola riga in corrispondenza di ogni blocco di X . In X^{cum} compaiono dunque solo righe di determinazioni delle variabili esplicative diverse tra loro. La k -esima riga della matrice X^{cum} è il vettore $\mathbf{x}'_k = (x_{k0}, \dots, x_{km})$.

Per $k = 1, \dots, K$, poniamo

$$Y_k^{cum} = \sum_{i=1}^{n_k} \frac{\omega_{ki}}{\omega_k} Y_{ki}, \quad \text{con } \omega_k = \sum_{i=1}^{n_k} \omega_{ki}.$$

Poiché i numeri aleatori Y_{ki} sono stocasticamente indipendenti, anche $Y_1^{cum}, \dots, Y_K^{cum}$ lo sono. Inoltre, Y_k^{cum} è media ponderata di numeri aleatori che soddisfano le ipotesi del Teorema 5.8.1, in particolare si noti che le loro distribuzioni hanno il medesimo valore del parametro canonico. Si ha allora

$$Y_k^{cum} \sim \exp\left\{ \frac{\omega_k}{\phi} [y\theta_k - b(\theta_k)] \right\} c(y, \phi, \omega_k),$$

con

$$E(Y_k^{cum}) = b'(\theta_k) = \mu_k = g^{-1}(\mathbf{x}'_k \boldsymbol{\beta}),$$

dove g è la funzione di collegamento del modello con dati individuali. Si ottiene un GLM per le medie ponderate Y_k^{cum} , $k = 1, \dots, K$, che diciamo *modello con dati raggruppati, associato al modello con dati individuali*. Nei due modelli sono comuni: la famiglia esponenziale lineare in cui sono scelte le distribuzioni delle variabili risposta, il parametro di dispersione, la struttura di regressione e la funzione di collegamento.

Nelle precedenti ipotesi, i modelli con dati individuali e con dati raggruppati conducono alla stessa stima di massima verosimiglianza del vettore $\boldsymbol{\beta}$ dei parametri di regressione. Lo proviamo mostrando che per i due modelli i sistemi delle equazioni di verosimiglianza sono equivalenti.

Le equazioni di verosimiglianza (5.4.4) per il modello con dati individuali sono

$$\sum_{k=1}^K \sum_{i=1}^{n_k} x_{kij} \frac{\omega_{ki}}{\phi} (y_{ki} - \mu_{ki}) \frac{1}{g'(\mu_{ki}) V(\mu_{ki})} = 0, \quad j = 0, \dots, m, \quad (5.8.1)$$

dove $\mu_{ki} = E(Y_{ki}) = \mu_k$ e $x_{kij} = x_{kj}$, $j = 0, \dots, m$, per $i = 1, \dots, n_k$. Allora, le equazioni del precedente sistema si possono riscrivere

$$\sum_{k=1}^K x_{kj} \frac{1}{\phi g'(\mu_k) V(\mu_k)} \sum_{i=1}^{n_k} \omega_{ki} (y_{ki} - \mu_k) = 0, \quad j = 0, \dots, m. \quad (5.8.2)$$

Posto $y_k^{cum} = \sum_{i=1}^{n_k} \frac{\omega_{ki}}{\omega_k} y_{ki}$, le equazioni diventano

$$\sum_{k=1}^K x_{kj} \frac{\omega_k}{\phi} (y_k^{cum} - \mu_k) \frac{1}{g'(\mu_k) V(\mu_k)} = 0, \quad j = 0, \dots, m. \quad (5.8.3)$$

Pertanto, i due sistemi (5.8.1), (5.8.3) sono equivalenti e le (5.8.3) sono le equazioni di verosimiglianza per il modello con dati raggruppati.

Osserviamo che se, in particolare, $\omega_{ki} = 1$ per ogni i , allora $\omega_k = n_k$, numerosità delle osservazioni nella classe k , e Y_k^{cum} è media aritmetica di Y_{ki} , $i = 1, \dots, n_k$.

Sottolineiamo che un modello con dati individuali e l'associato con dati raggruppati sono equivalenti ai fini di ottenere la stima dei parametri di regressione, ma non conducono alla stessa stima del parametro di dispersione (v. il prossimo Esempio 5.8.2 e § 6.3). Inoltre, come vedremo, anche i valori delle statistiche utilizzate per l'inferenza sono, in generale, diversi per i due modelli.

A partire da un modello con dati individuali è sempre possibile considerare un modello con dati raggruppati. Ciò può portare a considerevoli riduzioni della dimensione del problema se le variabili esplicative sono di classificazione, in quanto si passa da una matrice di regressione con $n_1 + \dots + n_K$ righe ad una con K righe. In presenza di variabili esplicative numeriche può invece accadere che le relative determinazioni siano comuni solo a poche osservazioni: non si ha quindi molta convenienza a raggruppare.

In presenza di variabili esplicative di classificazione è consigliato trattare un modello con dati raggruppati anziché con dati individuali. Infatti, alcune delle metodologie statistiche che si utilizzano per valutare la bontà di adattamento di un modello (v. in seguito § 6.2) sono affidabili solo nel caso di dati raggruppati e con elevata numerosità di osservazioni in ciascuna classe. Inoltre, da un punto di vista computazionale, la conseguente riduzione della dimensione del problema porta a riduzioni dei tempi di calcolo e di utilizzo di memoria.

Esempio 5.8.1. Modello per i numeri annui di sinistri con dati raggruppati

Riprendiamo l'Esempio 5.7.1 in cui abbiamo considerato un GLM con dati individuali per il numero annuo di sinistri. Al fine di costruire il GLM associato con dati raggruppati, in linea con le notazioni introdotte in questo paragrafo, modificiamo gli indici delle variabili risposta del modello con dati individuali, mettendo in evidenza anche la classe tariffaria di appartenenza. Sia

M_{ki} il numero annuo di sinistri per l' i -esimo assicurato in classe k .

Al numero aleatorio M_{ki} è assegnata distribuzione di Poisson e la funzione di collegamento è il logaritmo, si ha allora $E(M_{ki}) = \exp(x'_k \beta) = \mu_k$ e

$$Pr(M_{ki} = y) = \exp\{y\theta_{ki} - e^{\theta_{ki}}\} c(y) = \exp\{y\theta_k - e^{\theta_k}\} c(y)$$

dove $\theta_{ki} = \log(\mu_k) = \theta_k$.

Sono soddisfatte le ipotesi per la costruzione di un modello con dati raggruppati, associato al modello con dati individuali. Infatti il parametro canonico non dipende dall'osservazione, ma solo dalla classe tariffaria.

Poiché i pesi sono unitari, nel modello con dati raggruppati la variabile risposta, per ogni classe tariffaria, è la media aritmetica dei numeri di sinistri individuali e quindi è la frequenza sinistri nella classe

$$Y_k^{cum} = \sum_{i=1}^{n_k} \frac{1}{n_k} M_{ki} = \frac{1}{n_k} \sum_{i=1}^{n_k} M_{ki},$$

dove n_k è il numero di assicurati nella classe tariffaria k .

Per quanto visto in generale, si ha

$$Y_k^{cum} \sim \exp\left\{n_k(y\theta_k - e^{\theta_k})\right\} c(y, n_k).$$

Pertanto, Y_k^{cum} ha distribuzione di Poisson con peso n_k e

$$E(Y_k^{cum}) = e^{\theta_k} = \mu_k, \quad \text{var}(Y_k^{cum}) = \frac{1}{n_k} V(\mu_k) = \frac{1}{n_k} \mu_k.$$

Per le frequenze sinistri $Y_1^{cum}, \dots, Y_K^{cum}$ si ottiene dunque un GLM con distribuzioni di Poisson con pesi n_1, \dots, n_K , matrice di regressione X^{cum} e funzione di collegamento il logaritmo.

Consideriamo la specificazione del modello in SAS. Il *data set* di nome *polcuml*, ottenuto da *polizze1* dell'Esempio 5.7.1, contiene, per ogni classe individuata dai livelli delle variabili esplicative età dell'assicurato (*eta*) e potenza fiscale del veicolo (*potf*), le determinazioni delle variabili: numero totale di sinistri osservati nella classe (*nsinc*); numero di polizze nella classe (*npolc*); frequenza sinistri nella classe (*freqc*), rapporto tra il numero di sinistri e l'esposizione nella classe. Il seguente programma fornisce la stima dei parametri.

```
proc genmod data = polcuml;
  class eta potf;
  model freqc = eta potf / dist = poisson;
  weight npolc;
run;
```

Il numero di osservazioni è 12, pari al numero delle classi e il rango della matrice di regressione è 6. La Tabella 5.11 riporta l'*output* della procedura. Le stime dei parametri e dei numeri attesi coincidono con quelle dell'Esempio 5.7.1, sono invece diversi nei due casi i valori delle statistiche che forniscono indicazioni sulla bontà di adattamento. ♦

Esempio 5.8.2. Modello per i danni per sinistro con dati raggruppati

Riprendiamo l'Esempio 5.7.2. Anche in questo caso, al fine di costruire il modello associato con dati raggruppati, mettiamo in evidenza nelle variabili risposta del modello individuale anche la classe tariffaria di appartenenza. Sia

Y_{ki} l'importo del danno provocato dall' i -esimo sinistro in classe k .

Al numero aleatorio Y_{ki} è assegnata distribuzione gamma e la funzione di collegamento è il logaritmo, si ha allora $E(Y_{ki}) = \exp(\mathbf{x}'_k \boldsymbol{\beta}) = \mu_k$ e

$$Y_{ki} \sim \exp\left\{\frac{1}{\phi}(y\theta_{ki} + \log(-\theta_{ki}))\right\} c(y, \phi) = \exp\left\{\frac{1}{\phi}(y\theta_k + \log(-\theta_k))\right\} c(y, \phi),$$

con $\theta_{ki} = -1/\mu_k = \theta_k$.

$$\theta_{hi} = \bar{b}_{(hi)}^{(1)}$$

Poiché il parametro canonico dipende solo dalla classe tariffaria, sono soddisfatte le ipotesi per la costruzione di un modello associato con dati raggruppati.

I pesi sono unitari, pertanto nel modello con dati raggruppati la variabile risposta, per ogni classe tariffaria, è la media aritmetica dei danni per sinistro e quindi è il *danno medio per sinistro nella classe*

$$Y_k^{cum} = \sum_{i=1}^{m_k} \frac{1}{m_k} Y_{ki} = \frac{1}{m_k} \sum_{i=1}^{m_k} Y_{ki},$$

dove m_k è il numero di sinistri nella classe tariffaria k .

Per quanto visto in generale, si ha

$$Y_k^{cum} \sim \exp\left\{\frac{m_k}{\phi}(y\theta_k + \log(-\theta_k))\right\} c(y, \phi, m_k).$$

Tabella 5.11. Risultati del modello per i numeri di sinistri con dati raggruppati

The GENMOD Procedure									
Model Information									
Data Set					WORK.POLCUM1				
Distribution					Poisson				
Link Function					Log				
Dependent Variable					freqc				
Scale Weight Variable					npolc				
Observations Used					12				
Class Level Information									
Class	Levels	Values							
eta _a	4	18-22	23-26	27-43	>43				
potf	3	8-12	13-17	>17					
Criteria For Assessing Goodness Of Fit									
Criterion	DF	Value		Value/DF					
Deviance	6	7.1474		1.1912					
Scaled Deviance	6	7.1474		1.1912					
Pearson Chi-Square	6	7.0223		1.1704					
Scaled Pearson X2	6	7.0223		1.1704					
Log Likelihood		-23461.5448							
Algorithm converged.									
Analysis Of Parameter Estimates									
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq			
Intercept	1	-2.2802	0.0293	-2.3377 -2.2227	6041.05	<.0001			
eta _a 18-22	1	0.4489	0.0505	0.3498 0.5479	78.93	<.0001			
eta _a 23-26	1	0.2101	0.0409	0.1299 0.2902	26.39	<.0001			
eta _a 27-43	1	-0.1385	0.0267	-0.1908 -0.0862	26.92	<.0001			
eta _a >43	0	0.0000	0.0000	0.0000 0.0000	.	.			
potf 8-12	1	-0.2479	0.0350	-0.3165 -0.1793	50.14	<.0001			
potf 13-17	1	-0.1157	0.0314	-0.1773 -0.0541	13.56	0.0002			
potf >17	0	0.0000	0.0000	0.0000 0.0000	.	.			
Scale	0	1.0000	0.0000	1.0000 1.0000					

NOTE: The scale parameter was held fixed.

Pertanto, Y_k^{cum} ha distribuzione gamma con peso m_k e

$$E(Y_k^{cum}) = -\frac{1}{\theta_k} = \mu_k, \quad var(Y_k^{cum}) = \frac{\phi}{m_k} V(\mu_k) = \frac{\phi}{m_k} \mu_k^2.$$

Per i dati medi $Y_1^{cum}, \dots, Y_K^{cum}$ si ha dunque un GLM con distribuzioni gamma con pesi m_1, \dots, m_K , matrice di regressione X^{cum} e funzione di collegamento il logaritmo.

Consideriamo la specificazione del modello in SAS. A partire dal *data set* di nome *danni* dell'Esempio 5.7.2 si deve costruirne uno con dati raggruppati, che chiamiamo *dannicum*. Per ogni classe tariffaria individuata dai livelli delle variabili esplicative età dell'assicurato (*eta*) e potenza fiscale del veicolo (*potf*), il *data set* riporta le determinazioni delle variabili: numero totale di sinistri nella classe (*nsinc*); danno totale nella classe (*dannoc*); danno medio per sinistro nella classe (*dannomc*), rapporto tra il danno totale e il numero di sinistri nella classe. Le osservazioni sono dunque le classi tariffarie con polizze sinistrate.

Tabella 5.12. Risultati del modello per i dati per sinistro con dati raggruppati

The GENMOD Procedure									
Model Information									
Data Set					WORK.DANNICUM				
Distribution					Gamma				
Link Function					Log				
Dependent Variable					dannomc				
Scale Weight Variable					nsinc				
Observations Used					12				
Class Level Information									
Class	Levels	Values							
eta	4	18-22	23-26	27-43	>43				
potf	3	8-12	13-17	>17					
Criteria For Assessing Goodness Of Fit									
Criterion	DF	Value		Value/DF					
Deviance	6	122.4630		20.4105					
Scaled Deviance	6	12.0411		2.0069					
Pearson Chi-Square	6	145.7160		24.2860					
Scaled Pearson X2	6	14.3275		2.3879					
Log Likelihood		-87.8277							
Algorithm converged.									
Analysis Of Parameter Estimates									
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square				
Intercept	1	8.0565	0.0664	7.9263 8.1866	14715.9				
eta 18-22	1	0.3231	0.1067	0.1139 0.5322	9.16				
eta 23-26	1	0.2228	0.0931	0.0402 0.4053	5.72				
eta 27-43	1	0.0933	0.0638	-0.0317 0.2183	2.14				
eta >43	0	0.0000	0.0000	0.0000 0.0000	0.1436				
potf 8-12	1	-0.3504	0.0822	-0.5116 -0.1892	18.15				
potf 13-17	1	-0.1313	0.0700	-0.2684 0.0059	3.52				
potf >17	0	0.0000	0.0000	0.0000 0.0000	0.0606				
Scale	1	0.0983	0.0400	0.0443 0.2183	.				

NOTE: The scale parameter was estimated by maximum likelihood.

Il seguente programma fornisce la stima dei parametri.

```
proc genmod data = dannicum;
  class eta potf;
  model dannomc = eta potf / dist = gamma
    link = log;
  weight nsinc;
run;
```

Il numero di osservazioni usate è 12, pari al numero delle classi, e il rango della matrice di regressione è 6. La Tabella 5.12 riporta l'*output* della procedura.

Le stime dei parametri di regressione e dei dati attesi coincidono con quelle ottenute con il modello con dati individuali. Sono diversi, nei due modelli, i valori delle statistiche che forniscono indicazioni sulla bontà di adattamento. È inoltre diversa la stima del parametro di scala che nel modello con dati raggruppati è 0,0983. La stima del parametro di dispersione è ora $\hat{\phi} = 10,1729$, mentre nel modello con dati individuali si ha $\hat{\phi} = 1,2022$. Poiché è diverso, nei due modelli, il valore stimato di ϕ , sono diversi anche gli elementi che consentono di valutare le stime dei singoli parametri, in quanto tali elementi dipendono da ϕ (v. § 6.4 e § 6.6).

Le due stime di massima verosimiglianza del parametro di dispersione sono sensibilmente diverse, per indagare quale tra le due rispecchi in modo più aderente la dispersione dei dati, consideriamo la seguente analisi (v. Brockman, Wright (1992)). L'ipotesi probabilistica accolta per Y_{ki} comporta che

$$E(Y_{ki}) = \mu_k, \quad \text{var}(Y_{ki}) = \phi\mu_k^2.$$

Allora, posto $Y_{ki}^S = Y_{ki}/\mu_k$, si ha

$$E(Y_{ki}^S) = 1, \quad \text{var}(Y_{ki}^S) = \phi.$$

Una stima empirica di ϕ può essere ottenuta considerando la varianza campionaria dei rapporti $y_{ki}/\hat{\mu}_k$, dove y_{ki} è il valore osservato di Y_{ki} e $\hat{\mu}_k$ è la stima di μ_k ottenuta, indifferentemente, con uno dei due GLM. Con i dati dell'esempio, tale varianza campionaria è pari a 10,3061, un valore molto vicino alla stima di ϕ fornita dal modello con dati raggruppati. Nell'esempio, dunque, il modello con dati individuali e il metodo massima verosimiglianza portano a sottostimare la dispersione. Osserviamo però che il precedente confronto è effettuato ipotizzando che le variabili risposta abbiano distribuzioni gamma. Prima di trarre conclusioni, occorrerebbe analizzare l'adeguatezza di tale ipotesi in relazione ai dati, sfruttando le tecniche di controllo di un modello descritte nel Capitolo 6. ♦

5.9 Modelli con quasi-verosimiglianza

La classe dei GLM e le relative metodologie per l'inferenza statistica sono state estese e modificate in diversi modi per aumentarne l'ambito di applicabilità e la flessibilità. Una di tali estensioni porta ai *modelli con quasi-verosimiglianza*. Si tratta di modelli semiparametrici, nei quali si specificano solamente le strutture dei primi due momenti delle distribuzioni delle variabili risposta e non anche una particolare forma di distribuzione.

Per introdurre tali GLM semiparametrici, ricordiamo che una delle ipotesi alla base dei GLM è che le variabili risposta abbiano distribuzioni di probabilità appartenenti ad una famiglia esponenziale lineare. Come è stato osservato nel § 5.3, per le distribuzioni di tale classe, la scelta di una struttura per la speranza matematica, $\mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$, induce una struttura anche per la varianza, dato che si ha $\text{var}(Y_i) = \phi/\omega_i V(\mu_i) = \phi/\omega_i V(g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}))$.

I modelli con quasi-verosimiglianza consentono di svincolarsi dall'ipotesi che le variabili risposta abbiano distribuzioni appartenenti ad una famiglia esponenziale lineare, richiedendo solo la specificazione della struttura della speranza matematica e della collegata struttura della varianza. Precisamente, si considerano le seguenti ipotesi.

- Ipotesi probabilistiche. Le variabili risposta Y_1, \dots, Y_n sono stocasticamente indipendenti e, posto $E(Y_i) = \mu_i$, si ha $\text{var}(Y_i) = \phi/\omega_i V(\mu_i)$, $i = 1, \dots, n$, dove V è una funzione della speranza matematica, $\phi > 0$ è un parametro di dispersione e $\omega_i > 0$ è un peso assegnato.
- Ipotesi strutturali. Il legame tra la speranza matematica μ_i della variabile risposta Y_i ed il vettore \mathbf{x}_i delle determinazioni delle variabili esplicative, relativi all' i -esima unità statistica, è $\mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$, $i = 1, \dots, n$, con $\boldsymbol{\beta}$ vettore di parametri, g funzione di collegamento.

La V è ancora detta *funzione di varianza*, anche se non si fa qui riferimento alla classe esponenziale lineare.

Nelle precedenti ipotesi, si considera la funzione

$$Q(\boldsymbol{\mu}, \phi; \mathbf{y}) = \sum_{i=1}^n \omega_i \int_{y_i}^{\mu_i} \frac{y_i - s}{\phi V(s)} ds,$$

ammesso che esistano gli integrali che vi compaiono, che è detta *funzione di quasi-(log)-verosimiglianza*.

Assegnate la struttura di regressione e la funzione di collegamento, Q può essere vista come funzione di $\boldsymbol{\beta}$. Le derivate parziali di Q rispetto a β_0, \dots, β_m hanno la stessa forma delle derivate parziali (5.4.2) della log-verosimiglianza nei GLM, nelle quali in effetti compaiono solo i primi due momenti delle distribuzioni delle variabili risposta. Le derivate $\partial Q / \partial \beta_j$, $j = 0, \dots, m$, uguagliate a zero sono formalmente uguali alle equazioni (5.4.4), e sono dette *equazioni di Wedderburn o di quasi-verosimiglianza*:

$$\sum_{i=1}^n x_{ij} \frac{\omega_i}{\phi} (y_i - \mu_i) \frac{1}{g'(\mu_i) V(\mu_i)} = 0, \quad j = 0, \dots, m.$$

Il precedente è un sistema di equazioni di verosimiglianza solo se esiste una famiglia esponenziale lineare con funzione di varianza V e si decide di scegliere le distribuzioni delle variabili risposta in tale famiglia.

Le soluzioni del sistema delle equazioni di Wedderburn sono dette *stime di massima quasi-verosimiglianza* e si prova che soddisfano proprietà, come la consistenza e la normalità asintotica, analoghe a quelle delle stime di massima verosimiglianza.

Anche nei modelli con quasi-verosimiglianza, se le osservazioni possono essere ripartite in classi, ciascuna individuata da un medesimo vettore di determinazioni delle variabili esplicative, a partire da un modello con dati individuali si può costruirne uno

associato con dati raggruppati, equivalente ai fini della stima del vettore dei parametri di regressione β . Riprendendo le notazioni del § 5.8, con riferimento alla classe k , $k = 1, \dots, K$, siano x_k il vettore delle determinazioni delle variabili esplicative e, per l' i -esima osservazione nella classe, $i = 1, \dots, n_k$, siano Y_{ki} la variabile risposta, y_{ki} il valore osservato e ω_{ki} il peso. Supposto che le variabili risposta Y_{ki} , $k = 1, \dots, K$, $i = 1, \dots, n_k$, siano stocasticamente indipendenti e indicati con V la funzione di varianza, ϕ il parametro di dispersione, g la funzione di collegamento, nel modello con dati individuali si ha

$$E(Y_{ki}) = g^{-1}(x'_k \beta) = \mu_k, \quad \text{var}(Y_{ki}) = \frac{\phi}{\omega_{ki}} V(\mu_k).$$

Le equazioni di Wedderburn del modello sono formalmente uguali alle equazioni di verosimiglianza (5.8.2), relative al modello con dati individuali nel caso di distribuzione completamente specificata.

Con riferimento al modello sopra descritto, per $k = 1, \dots, K$, poniamo

$$Y_k^{cum} = \sum_{i=1}^{n_k} \frac{\omega_{ki}}{\omega_k} Y_{ki}, \quad \text{con } \omega_k = \sum_{i=1}^{n_k} \omega_{ki}.$$

I numeri aleatori $Y_1^{cum}, \dots, Y_K^{cum}$ sono stocasticamente indipendenti. Inoltre, si ha

$$E(Y_k^{cum}) = \sum_{i=1}^{n_k} \frac{\omega_{ki}}{\omega_k} E(Y_{ki}) = \mu_k = g^{-1}(x'_k \beta)$$

$$\text{e} \quad \text{var}(Y_k^{cum}) = \sum_{i=1}^{n_k} \frac{\omega_{ki}^2}{\omega_k^2} \text{var}(Y_{ki}) = \sum_{i=1}^{n_k} \frac{\omega_{ki}^2}{\omega_k^2} \frac{\phi}{\omega_{ki}} V(\mu_k) = \frac{\phi}{\omega_k} V(\mu_k).$$

Si ottiene un modello con quasi-verosimiglianza per le medie ponderate Y_k^{cum} , $k = 1, \dots, K$, che è detto *modello con dati raggruppati, associato al modello con dati individuali*. Nei due modelli sono comuni: la funzione di varianza, il parametro di dispersione, la struttura di regressione e la funzione di collegamento. Nella varianza della variabile risposta Y_k^{cum} , il peso è somma dei pesi ω_{ki} , $i = 1, \dots, n_k$. Le equazioni di Wedderburn del modello con dati raggruppati sono formalmente uguali alle equazioni di verosimiglianza (5.8.3), relative al modello con dati raggruppati nel caso di distribuzione completamente specificata. Poiché nel § 5.8 si è provato che i due sistemi di equazioni (5.8.2) e (5.8.3) sono equivalenti, il modello con quasi-verosimiglianza con variabili risposta $Y_1^{cum}, \dots, Y_K^{cum}$ è equivalente, ai fini della stima di β , al modello con dati individuali.

► Modelli con quasi-verosimiglianza in SAS

Per specificare in SAS un modello con quasi-verosimiglianza si deve indicare l'espressione della funzione di varianza e di una funzione detta *quasi-devianza* definita da

$$d(\mu, y) = -2 \sum_{i=1}^n \omega_i \int_{y_i}^{\mu_i} \frac{y_i - s}{V(s)} ds. \quad (5.9.1)$$

Poiché si ha

$$Q(\mu, \phi; y) = -\frac{1}{2\phi} d(\mu, y), \quad (5.9.2)$$

assegnata la funzione di quasi-devianza, si ottiene la funzione di quasi-verosimiglianza. Per qualche ulteriore commento sulla quasi-devianza si veda anche, al § 6.2, la nozione di devianza in un GLM e la sua espressione in funzione delle speranze matematiche delle variabili risposta.

Osserviamo che per calcolare la quasi-devianza è sufficiente conoscere la speranza matematica e la funzione di varianza delle distribuzioni delle variabili risposta.

Le istruzioni della procedura genmod sono le seguenti.

```
proc genmod data = archivio;
  class C1...Cm;
  variance var = espressione di V in funzione di _MEAN_;
  deviance dev = espressione di d in funzione di _MEAN_ e _RESP_;
  model risposta = X1...Xn C1...Cm / link = parola chiave
    <opzioni>;
  weight peso;
run;
```

Nel precedente programma, _MEAN_ e _RESP_ indicano variabili automaticamente definite che rappresentano rispettivamente la speranza matematica della variabile risposta e la variabile risposta, var e dev sono nomi di variabili che identificano le funzioni di varianza e di devianza. L'espressione della funzione d da indicare nell'istruzione deviance è la funzione che si ottiene considerando un generico addendo della (5.9.1), dopo avere calcolato l'integrale e sostituito _MEAN_ a μ_i , _RESP_ a y_i e posto peso ω_i unitario. I pesi sono dichiarati nell'istruzione weight.

Il parametro di scala del modello, la cui stima è riportata nell'*output* della procedura, è $\sqrt{\phi}$.

Alcuni esempi di modelli con quasi-verosimiglianza sono presentati nel Capitolo 8.

Il procedimento è un sistema di elaborazioni di variazioni di questo tipo che si basa su una struttura che si può considerare analogia alla struttura di un software di elaborazione dei dati. La struttura di base è quella di un insieme di "comandi" che possono essere eseguiti in sequenza.

Le strutture del sistema delle elaborazioni SAS sono state studiate in modo da consentire una grande libertà di scelta e di possibilità di elaborazione dei dati. La struttura di base è la struttura di base, analoga a quella delle strutture di elaborazione dei dati.

Anche nei modelli con quasi-verosimiglianza, la struttura di procedura SAS è quella di base, cioè una struttura di elaborazione dei dati con la possibilità di estensione e di personalizzazione delle strutture esistenti, a partire da un insieme di dati minimi e più semplici.