Sistema statistico nazionale

Istituto nazionale di statistica

Riponderazione

Note metodologiche

2005

A cura di: Stefano Falorsi.

Indice

1. STIMATORI DI PONDERAZIONE VINCOLATA

1.1. Premessa

Il presente capitolo è finalizzato ad illustrare le principali caratteristiche logiche ed algebriche degli stimatori di *ponderazione vincolata* (*calibration estimators* nella letteratura in lingua anglosassone sull'argomento, Deville e Särndal, 1992 §). Tale classe di stimatori si fonda sull'utilizzazione di variabili ausiliarie, per le quali sono noti i totali riferiti alla popolazione oggetto d'indagine. In tale contesto, i *pesi finali* si ottengono come soluzione di un problema di *minimo vincolato* che può essere così schematizzato:

(i) le *incognite* da determinare sono i pesi finali;

(ii) il *sistema di vincoli* assicura il rispetto della condizione di uguaglianza tra i totali noti (delle variabili ausiliarie) e le corrispondenti stime campionarie, calcolate sulla base dei pesi finali;

(iii) la *funzione obbiettivo* è una funzione di distanza tra i pesi finali incogniti e i *pesi diretti*, ottenuti come reciproco delle probabilità d'inclusione delle unità campionarie.

In sostanza la soluzione del suddetto problema di minimo vincolato conduce ad un insieme di pesi finali che consente di rispettare il sistema di vincoli e che contemporaneamente modifica il *meno possibile*, sulla base della funzione di distanza prescelta, l'insieme dei pesi diretti.

Tutti gli stimatori adottati nella pratica delle indagini campionarie su larga scala sono casi particolari dello stimatore di ponderazione vincolata, ottenuti definendo in modo opportuno la funzione di distanza impiegata. Nel caso in cui si utilizza la funzione di *distanza euclidea*, lo stimatore di ponderazione vincolata che si ottiene è uguale a quello di regressione generalizzata. Per quanto concerne gli altri stimatori di ponderazione vincolata, usualmente utilizzati nella pratica delle indagini campionarie (derivanti da funzioni di distanza per le quali sono verificate condizioni di regolarità piuttosto generali, Deville e Särndal, 1992 §), è possibile dimostrare la loro tendenza asintotica allo stimatore di regressione generalizzata. Quest'ultima caratteristica riveste una notevole importanza sia dal

punto di vista pratico che dal punto di vista teorico in quanto, nelle indagini su larga scala basate su campioni di notevole ampiezza, tutti gli stimatori di ponderazione vincolata assumono le medesime proprietà dello stimatore di regressione generalizzata, ossia:

- la correttezza asintotica;
- la correttezza dell'espressione linearizzata dello stimatore;
- l'esistenza di un'espressione esplicita della varianza e di uno stimatore consistente di tale varianza.

Una importante caratteristica dello stimatore di regressione generalizzata è quella di poter esplicitare il modello lineare, sottostante allo stimatore, che lega le variabili ausiliarie con le variabili di interesse. Tale possibilità è estesa anche a tutti gli stimatori ponderazione vincolata che convergono asintoticamente allo stimatore di regressione generalizzata. Infatti, per molti degli stimatori appartenenti alla suddetta classe non è possibile esplicitare chiaramente il modello sottostante ed ha, quindi, senso riferirsi al modello lineare del corrispondente stimatore di regressione generalizzata. Con riferimento ad un dato stimatore di regressione generalizzata, l'esplicitazione del modello lineare, implica necessariamente la definizione dei tre seguenti elementi fondamentali di: *gruppo di riferimento del modello*, *livello del modello* e *tipo di modello.* Nel *paragrafo 2*, che è dedicata alla descrizione degli stimatori di regressione generalizzata, viene svolta una trattazione approfondita dei suddetti elementi; nel presente paragrafo, tuttavia, essi vengono introdotti con riferimento alla classe più generale degli stimatori di ponderazione vincolata. I tre elementi in parola assumono comunque degli aspetti leggermente diversi da quelli relativi allo stimatore di regressione generalizzata; la principale differenza, in tale contesto, consiste nel fatto che per la loro definizione si deve far riferimento al problema di minimo vincolato sottostante alla costruzione degli stimatori di ponderazione vincolata; in tale ottica, sembra più logico parlare di: (i) *gruppo di riferimento dello stimatore*, (ii) *livello dello stimatore*, e (iii) *tipo dello stimatore*; questi ultimi concetti saranno brevemente illustrati nel seguito.

Si dice *gruppo di riferimento dello stimatore* un sottoinsieme (o sottopopolazione) della popolazione oggetto d'indagine con riferimento al quale:

- sono noti i totali della popolazione di una o più variabili ausiliarie;
- viene definito il problema di minimo vincolato sottostante lo stimatore.

I *gruppi* rappresentano una partizione completa della popolazione.

Il concetto di *livello dello stimatore* è relativo al tipo di unità utilizzata nella formulazione del problema di minimo vincolato. Se le unità sulle quali è definito il problema sono costituite dai singoli elementi della popolazione, lo stimatore è definito al *livello di unità elementari*; in tal caso, le variabili di interesse e quelle ausiliarie si riferiscono ai singoli elementi della popolazione. Se invece, le unità su cui è definito il problema di minimo vincolato sono costituite da gruppi o *cluster* di singoli elementi della popolazione, lo stimatore è definito al *livello di cluster*, in tal caso le variabili di interesse e quelle ausiliarie, si riferiscono a *cluster* di elementi della popolazione.

Per quanto riguarda il concetto di *tipo di stimatore*, esso viene essenzialmente definito dal numero e dal tipo di variabili ausiliarie specificate nel problema di minimo vincolato.

1.2. Caratteristiche generali degli stimatori di ponderazione vincolata

Sia U una popolazione finita di N elementi, che indichiamo come $U = \{1,..., k,..., N\}$, e sia s un campione casuale di n elementi estratto da U, che indichiamo come $s = \{1,..., k,..., n\}$, mediante un disegno di campionamento che genera l'*universo dei campioni S* (ossia l'insieme di tutti i possibili campioni estraibili mediante il disegno in parola) ed assegna al generico campione s la probabilità p(s) di essere estratto, dove $\sum_{s \in S} p(s) = 1$. Con riferimento alla generica unità $k \in U$, indichiamo quindi con: $\pi_k = \sum_{s \in S(k)} p(s)$, la probabilità di inclusione nel campione dell'unità, dove *S*(k) denota il sottoinsieme di *S* caratterizzato dai campioni contenenti l'unità in oggetto; y_k, il valore assunto dalla

variabile di interesse *y*; $\mathbf{x}_k = \left(x_{1k},\ldots,x_{jk},\ldots,x_{Jk}\right)$, il valore assunto dal vettore $\boldsymbol{x} = \left(x_1,\ldots,x_j,\ldots,x_J\right)'$ di J variabili ausiliarie.

Si vuole stimare il totale Y della variabile *y* , dato dalla seguente espressione:

$$Y = \sum_{k\in U} y_k \qquad (1)$$

sulla base delle seguenti informazioni:

– per ciascun elemento del campione s si dispone delle J+1 osservazioni $\left(y_k, \mathbf{x}_k\right)$;

– risultano conosciuti i J valori del vettore $\mathbf{X} = (X_1,\ldots,X_j,\ldots,X_J)'$ dei totali delle J variabili ausiliarie, in cui

$$X_j = \sum_{k\in U} x_{jk} \qquad (j=1,\ldots,J). \qquad (2)$$

Uno stimatore del totale Y appartenente alla classe degli stimatori di ponderazione vincolata, può essere espresso mediante la seguente relazione

$$\boxed{\tilde{Y}_{PV} = \sum_{k\in s} y_k d_k \gamma_k = \sum_{k\in s} y_k w_k}, \qquad (3)$$

in cui: $d_k = \pi_k^{-1}$ (per k=1,.....,n) indica il *peso diretto*, $w_k = d_k \gamma_k$ denota il *peso finale* associato a tale unità, essendo γ_k il correttore del peso diretto.

L'insieme dei pesi finali $\left\{w_k;\ k = 1,\ldots.,n\right\}$ è ottenuto come soluzione di un problema di minimo vincolato in cui la funzione obbiettivo è data da[1]

[1] La (4) è equivalente a minimizzare il valore atteso sotto il disegno di campionamento della funzione obiettivo, in quanto la (4) è valida per ciascun campione s estratto nello spazio campionario.

$$\min\left\{\sum_{k\in s} G_k\left(w_k; d_k\right)\right\} \qquad (4)$$

ed i vincoli sono espressi dal sistema di J equazioni

$$\sum_{s\in k} w_k \mathbf{x}_k = \mathbf{X}, \qquad (5)$$

dove con $G_k\left(w_k; d_k\right)$ si è indicata una funzione di distanza tra il peso diretto d_k ed il peso finale w_k, ovvero una funzione definita sulla variabile w_k in cui d_k rappresenta una costante nota (o *parametro*) della funzione stessa. Il nostro obbiettivo è, quindi, quello di individuare un insieme di pesi finali $\left\{w_k;\ k=1,\ldots,n\right\}$ che consenta di rispettare il sistema di vincoli (5) e che contemporaneamente modifichi il *meno possibile* , sulla base della funzione di distanza prescelta, l'insieme dei pesi diretti $\left\{d_k; k=1,\ldots,n\right\}$.

Affinché il problema di minimo vincolato, definito dalla (4) e dalla (5), ammetta una soluzione e che tale soluzione sia unica la funzione di distanza $G_k\left(w_k; d_k\right)$ deve soddisfare le seguenti condizioni di regolarità:

(a) per ogni fissato $d_k > 0$ esiste un'intervallo $I(d_k)$, contenente d_k, in cui $G_k\left(w_k; d_k\right)$ sia strettamente convessa, differenziabile rispetto a w_k e non negativa, e si abbia inoltre $G_k\left(d_k; d_k\right) = 0$;

(b) la derivata prima di $G_k\left(w_k; d_k\right)$, indicata con $g_k\left(w_k; d_k\right) = \dfrac{\delta G_k\left(w_k; d_k\right)}{\delta w_k}$, deve essere una funzione continua e biunivoca nell'intervallo $I(d_k)$. Da ciò deriva che $g_k\left(w_k; d_k\right)$ nell'intervallo $I(d_k)$ è una funzione strettamente crescente di w_k ed inoltre $g_k\left(d_k, d_k\right) = 0$. Inoltre, poiché

la funzione è strettamente crescente e continua, esiste la funzione inversa, $g_k^{-1}(\cdot)$, ovvero una funzione per la quale vale

$$w_k = g_k^{-1}\left(g_k\left(w_k; d_k\right)\right). \qquad (6)$$

Al fine di ottenere il vettore **w =** $(w_1,..., w_k..., w_n)'$ soluzione del problema di minimo vincolato (4) e (5), si definisce la seguente funzione di Lagrange

$$L(\boldsymbol{\lambda}, \mathbf{w}) = \sum_{k \in s} G_k\left(d_k, w_k\right) - \left(\sum_{k \in s} w_k \mathbf{x}_k - \mathbf{X}\right)' \boldsymbol{\lambda}$$

in cui $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_j,\ldots,\lambda_J)'$ è il vettore dei moltiplicatori di Lagrange. Si risolve, quindi, il sistema omogeneo di (n + J) equazioni nelle (n+J) incognite (**w**, $\boldsymbol{\lambda}$)

$$\begin{cases} \dfrac{\delta L(\boldsymbol{\lambda}, \mathbf{w})}{\delta w_k} = g_k\left(w_k; d_k\right) - \mathbf{x}_k' \boldsymbol{\lambda} = 0 & \text{per } k = 1,\ldots,n \\ \dfrac{\delta L(\boldsymbol{\lambda}, \mathbf{w})}{\delta \lambda_j} = \sum_{k \in s} w_k x_{jk} - X_j = 0 & \text{per } j = 1,\ldots,J \end{cases} \qquad (7)$$

Sulla base della (6) è possibile scrivere le prime n equazioni del sistema (7) nel seguente modo

$$g_k^{-1}(g_k(w_k; d_k)) = g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right),$$

da cui deriva

$$w_k = g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right). \qquad (8)$$

Poiché, come risulta dalla relazione (3), l'obbiettivo è quello di ottenere un'espressione del peso finale come prodotto del peso diretto per un coefficiente di correzione, è possibile sulla base della (8) individuare il coefficiente in oggetto mediante i seguenti semplici passaggi

$$w_k = d_k \, \gamma_k = g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right),$$

da cui deriva

$$\gamma_k = \frac{1}{d_k} g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right).$$

Ponendo quindi $F_k\left(\mathbf{x}_k'\boldsymbol{\lambda}\right) = \frac{1}{d_k} \; g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$, si ottiene ovviamente

$$\boxed{w_k = d_k \, F_k\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)} \qquad (9)$$

L'espressione (9) assume una notevole importanza in quanto da essa si desume che il peso finale w_k della generica unità k-esima (k=1,...,n) si ottiene moltiplicando il peso diretto d_k per un coefficiente di correzione scalare $F_k\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$, funzione della variabile $u_k = \left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$ combinazione lineare del vettore di variabili ausiliarie $\mathbf{x}_k$ e dei J valori incogniti del vettore $\boldsymbol{\lambda}$.

La (9) non è ancora una relazione operativa in quanto non sono noti i valori numerici del vettore $\boldsymbol{\lambda}$; a tal fine introduciamo la (9) nelle ultime J equazioni del sistema (7) ottenendo in tal modo un sistema di J equazioni nelle J incognite $\lambda_1,\ldots,\lambda_j,\ldots,\lambda_J$

$$\sum_{k \in s} d_k \, x_{jk} \, F_k(\mathbf{x}_k'\boldsymbol{\lambda}) = X_j \quad \text{per} \;\; (j=1,\ldots,J),$$

esprimibile in termini vettoriali come

$$\sum_{k \in s} d_{ks} \mathbf{x}_k F_k(\mathbf{x}_k' \boldsymbol{\lambda}) = \mathbf{X}, \qquad (10)$$

che è equivalente a

$$\mathbf{X} - \sum_{k \in s} d_k \mathbf{x}_k = \sum_{k \in s} d_k \mathbf{x}_k F_k(\mathbf{x}_k' \boldsymbol{\lambda}) - \sum_{k \in s} d_k \mathbf{x}_k =$$

$$= \sum_{k \in s} d_k \mathbf{x}_k \left(F_k(\mathbf{x}_k' \boldsymbol{\lambda}) - 1 \right). \qquad (11)$$

.

Infine, indicando con:

$$\boldsymbol{\phi}(\boldsymbol{\lambda}) = \sum_{k \in s} d_k \mathbf{x}_k \left(F_k(\mathbf{x}_k' \boldsymbol{\lambda}) - 1 \right), \qquad (12)$$

è possibile riscrivere il sistema (11) nel seguente modo

$$\boldsymbol{\phi}(\boldsymbol{\lambda}) = \mathbf{X} - \sum_{k \in s} d_k \mathbf{x}_k \qquad (13)$$

la cui j-esima (j=1,...,J) equazione, che indichiamo $\phi_j(\boldsymbol{\lambda})$, è data da

$$\sum_{k \in s} d_k x_{jk} \left(F_k(\mathbf{x}_k' \boldsymbol{\lambda}) - 1 \right) = X_j - \sum_{k \in s} d_k x_{jk} . \qquad (14)$$

Una soluzione numerica[2] al sistema (13) può essere ottenuta in modo iterativo mediante il metodo di Newton. Per illustrare tale metodo, indichiamo con ν (ν=1,2,....) la generica iterazione e con $\boldsymbol{\lambda}_\upsilon = (\lambda_{1,\nu},\ldots,\lambda_{j,\nu},\ldots,\lambda_{J,\nu})'$ i valori di $\boldsymbol{\lambda}$ relativi all'iterazione. I passi dell'algoritmo sono i seguenti:

1. si pone il valore iniziale di $\boldsymbol{\lambda}$, che indichiamo come $\boldsymbol{\lambda}_0$, pari a

   $\boldsymbol{\lambda}_0 = \mathbf{0}$,

   dove **0** indica un vettore di dimensione J i cui elementi sono tutti pari a zero;

2. i valori $\boldsymbol{\lambda}_\upsilon$ alle successive iterazioni (ν =1,2,...) sono dati da:

$$\boldsymbol{\lambda}_\nu = \boldsymbol{\lambda}_{\nu-1} + \left\{ \left[\frac{\delta \boldsymbol{\phi}(\boldsymbol{\lambda})}{\delta \boldsymbol{\lambda}} \right]_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{\nu-1}} \right\}^{-1} \left\{ \mathbf{X} - \tilde{\mathbf{X}} - \boldsymbol{\phi}(\boldsymbol{\lambda}_{\nu-1}) \right\}, \qquad (15)$$

in cui: $\tilde{\mathbf{X}} = \sum_{k \in s} d_k \, \mathbf{x}_k$; $\boldsymbol{\phi}(\boldsymbol{\lambda}_{\nu-1})$ è il vettore i cui J valori sono ottenuti ponendo nella (12) $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\nu-1}$; $\left\{ \left[\frac{\delta \boldsymbol{\phi}(\boldsymbol{\lambda})}{\delta \boldsymbol{\lambda}} \right]_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{\nu-1}} \right\}$ è una matrice simmetrica di dimensione (J x J), il cui generico elemento , $a_{ji}(\boldsymbol{\lambda}_{\nu-1})$ sulla riga j-esima e sulla colonna i-esima è la derivata prima di $\phi_j(\boldsymbol{\lambda})$ rispetto a λ_i, calcolata ponendo $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\upsilon-1}$;

3. Si itera il passo 2 finché non viene verificata almeno una delle due condizioni di seguito riportate:

$$\operatorname*{Max}_j \left(\frac{\left| \lambda_{j,\nu-1} - \lambda_{j,\nu} \right|}{\left| \lambda_{j,\nu-1} \right|} \right) \leq \omega \qquad (16)$$

$$\nu = \nu_{MAX}\,, \qquad (17)$$

---

[2] Altre soluzioni per specifiche funzioni di distanza sono riportate nel lavoro di Singh e Mohl (1996 §).

dove ω è una costante piccola a piacere scelta nell' intervallo (0 , 1), e ν_{MAX} indica il numero massimo di iterazioni ammesse, oltre il quale si giudica che l'algoritmo non converga. Mediante la (16) si interrompe il processo iterativo quando tra l'iterazione ν e l'iterazione precedente (ν-1) la maggiore differenza relativa sui valori dei λ_j (j=1,...,J) è minore di un valore piccolo a piacere. La condizione (17) viene introdotta al fine di interrompere le iterazioni quando il processo non converge.

A conclusione di questo paragrafo riteniamo utile riassumere i passi analitici ed operativi necessari alla costruzione di uno stimatore di ponderazione vincolata:

1. per ciascuna unità del campione si calcolano i pesi diretti, $d_k = \pi_k^{-1}$ (k=1,...,n),
2. si sceglie la funzione di distanza $G_k\left(w_{ks}; d_k\right)$;
3. si definisce la funzione $g_k\left(w_{ks}; d_k\right)$, ottenuta come derivata prima di $G_k\left(w_{ks}; d_k\right)$;
4. dall'equazione $g_k(w_{ks}; d_k) - \mathbf{x}_k'\boldsymbol{\lambda} = 0$ si determina la funzione inversa $w_{ks} = g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$;
5. si ottiene l'espressione funzionale $F_k\left(\mathbf{x}_k'\boldsymbol{\lambda}\right) = \frac{1}{d_k} g_k^{-1}\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$ del correttore γ_{ks} del peso diretto;
6. si determinano i valori di $\boldsymbol{\lambda}$, risolvendo il sistema $\boldsymbol{\phi}(\boldsymbol{\lambda}) = \mathbf{X} - \sum_{k \in s} d_k \mathbf{x}_k$, secondo il metodo di Newton;
7. si calcolano i valori numerici di correttori γ_{ks} (k=1,...,n) sostituendo i valori di $\boldsymbol{\lambda}$ nelle espressioni funzionali $F_k\left(\mathbf{x}_k'\boldsymbol{\lambda}\right)$;
8. si determinano i pesi finali mediante il prodotto $w_{ks} = d_k\, \gamma_{ks}$;

9. è possibile, quindi, calcolare lo stimatore di ponderazione vincolata $\tilde{Y}_{PV} = \sum_{k \in s} y_k w_{ks}$.

1.3. Scelta della funzione di distanza

In questo paragrafo vengono esplicitati i passi analitici ed operativi sopra descritti, necessari alla costruzione di uno stimatore di ponderazione vincolata, con riferimento alle più importanti funzioni di distanza utilizzate nelle indagini campionarie su larga scala. Nella seguente tabella vengono considerate alcune delle più importanti funzioni di distanza note nella letteratura specialistica sull'argomento.

Tabella 1 - Funzioni di distanza

Nome	Espressione
Lineare	$\frac{(w_k - d_k)^2}{d_k}$
Lineare troncata	$\begin{cases} \frac{(w_k - d_k)^2}{d_k} & \text{se} \quad L < \frac{w_k}{d_k} < U \\ \infty & \text{altrimenti} \end{cases}$
Esponenziale	$w_k \ln\left(\frac{w_k}{d_k}\right) - w_k + d_k$
Logit	$\left(\frac{w_k}{d_k} - L\right) \ln\left(\frac{\frac{w_k}{d_k} - L}{1 - L}\right) + \left(U - \frac{w_k}{d_k}\right) \ln\left(\frac{M - \frac{w_k}{d_k}}{U - 1}\right)$
Chi-quadrato modificato	$\frac{w_k}{2}\left(\frac{w_k}{d_k} - 1\right)^2$
Minima entropia	$-d_k \ln\left(\frac{w_k}{d_k}\right) + w_k - d_k$
Hellinger	$2d_k\left(\sqrt{\frac{w_k}{d_k}} - 1\right)^2$

Le funzioni di distanza appena considerate soddisfano le proprietà di regolarità descritte nel precedente paragrafo e sono tutte asintoticamente equivalenti alla funzione di distanza lineare.

Per quanto riguarda la scelta della funzione di distanza, nelle indagini su larga scala, essa è determinata sia dall'intervallo di variazione dei pesi finali che dall'esistenza di una soluzione, qualora il sistema dei vincoli sia congruente. A tale proposito vale la pena osservare che:

- l'utilizzazione della funzione di distanza *lineare* può condurre alla determinazione di alcuni pesi negativi, in quanto l'intervallo di variazione ammesso per i pesi finali è del tipo $(-\infty,+\infty)$;
- le funzioni di distanza *esponenziale*, *chi-quadrato modificato*, *minima entropia* e *hellinger*, garantiscono l'ottenimento di pesi finali tutti positivi;
- la funzione di distanza esponenziale può condure alla determinazione di pesi finali estremamente alti in confronto ai corrispondenti pesi diretti, in quanto l'intervallo di variazione ammesso per i pesi finali è $(0,+\infty)$;
- le funzioni *logit* e *lineare troncata* hanno l'importante proprietà di condurre alla determinazione di pesi finali tutti inclusi nell'intervallo $(\mathrm{Ld}_k, \mathrm{Ud}_k)$, dove L ed U sono parametri delle funzioni che possono essere specificati direttamente dall'utente. In tal modo, a differenza della funzione di distanza esponenziale, è possibile evitare l'ottenimento di pesi finali estremamente alti pur mantenendo le proprietà asintotiche di convergenza allo stimatore di regressione generalizzata;
- le funzioni di distanza lineare ed esponenziale, hanno l'importante proprietà di condurre certamente ad una soluzione qualora il sistema dei vincoli sia congruente.

Al fine di non appesantire la trattazione seguente verranno descritte in dettaglio solamente le funzioni di distanza lineare, esponenziale e logit troncata che si ritengono utili a risolvere la maggior parte dei problemi di stima che si

incontrano nelle indagini concrete su larga scala[3] (Singh e Mohl, 1996). Le ragioni di tale scelta risiedono nel fatto che tali funzioni rivestono un importanza particolare rispetto alle altre; infatti, le funzioni lineare ed esponenziale, a differenza delle altre, garantiscono l'ottenimento di una soluzione al sistema di minimo vincolato qualora il sistema dei vincoli sia congruente; inoltre la funzione di distanza logit permette di restringere il campo di variabilità dei pesi finali definendo opportunamente i parametri L ed U.

Distanza euclidea

La funzione di distanza è espressa da

$$G\left(w_{ks};\, d_k\right) = \frac{\left(d_k - w_{ks}\right)^2}{q_k d_k}\ , \qquad (18)$$

in cui $1/q_k$ indica un peso non correlato a d_k assegnato all'unità k-esima. Nella maggior parte delle applicazioni si utilizza il peso uniforme $1/q_k = 1$, ma in alcuni casi può essere conveniente utilizzare pesi $1/q_k$ variabili[4].

La funzione $g_k\left(w_{ks}; d_k\right)$, ottenuta come derivata prima della (18) è data da

$$g_k\left(w_{ks}; d_k\right) = -\frac{2}{q_k d_k}\left(d_k - w_{ks}\right). \qquad (19)$$

---

[3]Nei lavori di Deville e Särndal (1992) e di Singh e Mohl (1996 §) vengono introdotte altre funzioni di distanza. In questa sede limitiamo però l'esposizione unicamente i tre metodi adottati nelle indagini Istat ed utili a risolvere la maggior parte dei problemi di stima che si pongono nelle indagini su larga scala.

[4]Nel lavoro di Alexander (1987), si dimostra l'utilità dell' adozione dei pesi $1/q_k$ variabili, per risolvere particolari problemi di sottocopertura.

Sulla base della (19) è possibile scrivere le prime n equazioni del sistema (7) come

$$-\frac{2}{q_k d_k}(d_k - w_{ks}) = \mathbf{x}_k'\boldsymbol{\lambda}\,, \qquad \text{per k=1,....,n}$$

la cui soluzione esplicita è

$$w_{ks} = d_k(1+\frac{1}{2}q_k\mathbf{x}_k'\boldsymbol{\lambda}) = g_k^{-1}(\mathbf{x}_k'\boldsymbol{\lambda})\,,$$

da cui si evince che

$$F_k(\mathbf{x}_k'\boldsymbol{\lambda}) = (1+\frac{1}{2}q_k\mathbf{x}_k'\boldsymbol{\lambda}) \qquad (20)$$

Mediante le formule (12) e (13) si ottiene il sistema di J equazioni in J incognite

$$\sum_{k\in s}\frac{1}{2}q_k d_k \mathbf{x}_k\mathbf{x}_k'\boldsymbol{\lambda} = \mathbf{X} - \tilde{\tilde{\mathbf{X}}}$$

da cui si ha

$$\boldsymbol{\lambda} = \left(\sum_{k\in s}\frac{1}{2}q_k d_k \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\mathbf{X} - \tilde{\tilde{\mathbf{X}}}\right).$$

Introducendo l'espressione esplicita di $\boldsymbol{\lambda}$ nella (20) si ottiene

$$F_k(\mathbf{x}_k'\boldsymbol{\lambda}) = 1 + \frac{1}{2} q_k \mathbf{x}_k' \left(\sum_{k \in s} \frac{1}{2} q_k d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\mathbf{X} - \tilde{\mathbf{X}} \right)$$

che è equivalente a

$$F_k(\mathbf{x}_k'\boldsymbol{\lambda}) = 1 + (\mathbf{X} - \tilde{\mathbf{X}})' \left(\sum_{k \in s} \frac{1}{2} q_k d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \frac{1}{2} q_k d_k \mathbf{x}_k \qquad (21)$$

il generico peso finale w_{ks} (per k=1,....,n) viene infine determinato mediante la (9).

Distanza logaritmica

La funzione di distanza è espressa da

$$G\left(w_{ks}; d_k\right) = \frac{w_{ks}}{q_k} \ln\left(\frac{w_{ks}}{d_k}\right) - w_{ks} + d_k \ , \qquad (22)$$

in cui il simbolo “ln” indica il logaritmo naturale in base e.

La funzione $g_k\left(w_{ks}; d_k\right)$, ottenuta come derivata prima della (22) è data da

$$g_k(w_{ks}; d_k) = \frac{1}{q_k} \ln\left(\frac{w_{ks}}{d_k}\right). \qquad (23)$$

Sulla base della (23) è possibile scrivere le prime n equazioni del sistema (7) come

$$\frac{1}{q_k}\ln\left(\frac{w_{ks}}{d_k}\right) = \mathbf{x}_k'\boldsymbol{\lambda}\,, \qquad \text{per k=1,....,n}$$

la cui soluzione esplicita è

$$w_{ks} = d_k \exp(q_k \mathbf{x}_k'\boldsymbol{\lambda})\,,$$

da cui si evince che

$$F_k(\mathbf{x}_k'\boldsymbol{\lambda}) = \exp(q_k \mathbf{x}_k'\boldsymbol{\lambda}) \qquad (24)$$

Sostituendo la (24) nella (12) si ottiene

$$\boldsymbol{\phi}(\boldsymbol{\lambda}) = \sum_{k\in s} d_k \mathbf{x}_k \left(\exp(q_k \mathbf{x}_k'\boldsymbol{\lambda}) - 1\right), \qquad (25)$$

e quindi il sistema (13) di J equazioni in J incognite può essere riscritto come

$$\sum_{k\in s} d_k \mathbf{x}_k \left(\exp(q_k \mathbf{x}_k'\boldsymbol{\lambda}) - 1\right) = \mathbf{X} - \tilde{\mathbf{X}}\,. \qquad (26)$$

Tale sistema, di tipo non lineare, può essere risolto mediante il metodo di Newton descritto nella (15). Dalla (25) si ha quindi che:

$$\left[\frac{\delta\boldsymbol{\phi}(\boldsymbol{\lambda})}{\delta\boldsymbol{\lambda}}\right]_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{\nu-1}} = \sum_{k\in s} q_k d_k \mathbf{x}_k \mathbf{x}_k' \exp(q_k \mathbf{x}_k'\boldsymbol{\lambda}_{\nu-1})$$

il cui generico elemento generico elemento $a_{ji}(\boldsymbol{\lambda}_{\nu-1})$ sulla riga j-esima e sulla colonna i-esima della matrice è espresso come

$$a_{ji}(\boldsymbol{\lambda}_{\nu-1}) = \sum_{k \in s} q_k d_k x_{jk} x_{ik} \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}_{\nu-1}) .$$

Indichiamo, quindi, con $\boldsymbol{\lambda}_*$ il vettore di J valori numerici, soluzione del sistema (26), ottenuti mediante il metodo di Newton. Sostituendo i valori così ottenuti nell'espressione (24) è possibile calcolare il valore numerico $F_k(\mathbf{x}_k' \boldsymbol{\lambda}_*) = \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}_*)$ per ciascuna unità k=1,....,n. Sostituendo infine tali valori nella (9) è possibile calcolare l'insieme dei pesi finali w_{ks} (per k=1,....,n) .

Distanza logit o logaritmica limitata

La funzione di distanza è espressa da

$$G(w_{ks}; d_k) = \frac{d_k}{A q_k}\left(\frac{w_{ks}}{d_k} - L\right) \ln \frac{\frac{w_{ks}}{d_k} - L}{1 - L} + \frac{d_k}{A q_k}\left(U - \frac{w_{ks}}{d_k}\right) \ln \frac{U - \frac{w_{ks}}{d_k}}{U - 1} , \qquad (27)$$

dove L ed U sono due costanti tali che $L < 1 < U$ ed

$$A = \frac{(U - L)}{(U - 1)(1 - L)}$$

La funzione $g_k(w_{ks}; d_k)$, ottenuta come derivata prima della (27) è data da

$$g_k(w_{ks}; d_k) = \frac{1}{A q_k}\left\{ \ln\left(\frac{\frac{w_{ks}}{d_k} - L}{1 - L}\right) - \ln\left(\frac{U - \frac{w_{ks}}{d_k}}{U - 1}\right)\right\}. \qquad (28)$$

Sulla base della (28) è possibile scrivere le prime n equazioni del sistema (7) come

$$\frac{1}{Aq_k}\left\{\ln\left(\frac{\frac{w_{ks}}{d_k}-L}{1-L}\right)-\ln\left(\frac{U-\frac{w_{ks}}{d_k}}{U-1}\right)\right\}=\mathbf{x}_k'\boldsymbol{\lambda}\,,\qquad \text{per k=1,....,n}$$

la cui soluzione esplicita rispetto al peso finale è data da

$$w_{ks}=d_k\,\frac{L(U-1)+U(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}\,,$$

da cui si evince che

$$F_k(\mathbf{x}_k'\boldsymbol{\lambda})=\frac{L(U-1)+U(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}\qquad(29)$$

Sostituendo la (29) nella (12) si ottiene

$$\boldsymbol{\phi}(\boldsymbol{\lambda})=\sum_{k\in s}d_k\mathbf{x}_k\left(\frac{L(U-1)+U(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}_k\boldsymbol{\lambda})}-1\right),\qquad(30)$$

e quindi il sistema (13) di J equazioni in J incognite può essere riscritto come

$$\sum_{k\in s}d_k\mathbf{x}_k\exp(q_k\mathbf{x}_k'\boldsymbol{\lambda})=\mathbf{X}-\tilde{\mathbf{X}}.\qquad(26)$$

Tale sistema, di tipo non lineare, può essere risolto mediante il metodo di Newton descritto nella (15). Dalla (30) si ha quindi che:

$$\left[\frac{\delta\phi(\lambda)}{\delta\lambda}\right]_{\lambda=\lambda_{\nu-1}} = \sum_{k\in s} q_k d_k \mathbf{x}_k \mathbf{x}_k' \exp(q_k \mathbf{x}_k' \lambda_{\nu-1})$$

il cui elemento generico elemento $a_{ji}(\lambda_{\nu-1})$ sulla riga j-esima e sulla colonna i-esima della matrice è espresso come

$$a_{ji}(\lambda_{\nu-1}) = \sum_{k\in s} q_k d_k x_{jk} x_{ik} \exp(q_k \mathbf{x}_k' \lambda_{\nu-1}) .$$

Indichiamo, quindi, con $\lambda = \lambda_*$ il vettore di J valori numerici, soluzione del sistema (26), ottenuti mediante il metodo di Newton. Sostituendo i valori così ottenuti nell'espressione (24) è possibile calcolare il valore numerico $F_k(\mathbf{x}_k'\lambda) = \exp(q_k \mathbf{x}_k'\lambda)$ per ciascuna unità k (k=1,....,n). Sostituendo infine tali valori nella (9) è possibile calcolare l'insieme dei pesi finali w_{ks} (per k=1,....,n) .

Le funzioni di distanza euclidea e logaritmica troncata hanno la desiderabile proprietà di portare sempre ad una soluzione qualora il sistema dei vincoli sia congruente; si dimostra (Deville and Särndal, 1992, p.p 379) che, per campioni sufficientemente grandi, le suddette funzioni conducono a stimatori aventi approssimativamente la stessa varianza; pertanto al fine di pervenire ad una scelta tra di esse è necessario analizzare l'intervallo dei valori che i coefficienti di correzione $F_k(\mathbf{x}_k'\lambda)$ assumono nei due casi.

La (18) è la funzione di distanza che conduce allo stimatore di regressione generalizzato (Särndal, Swensson e Wretman, 1992; Isaki and Fuller, 1982). Lo stimatore in oggetto viene adottato per l'ottenimento delle stime dell'indagine canadese sulle Forze di Lavoro ed è stato applicato in ambito ISTAT per il calcolo delle stime dell'indagine sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari 1986-1987 (ISTAT , 1991). Essa porta a coefficienti di correzione che possono variare nell'intervallo (-∞ , ∞) e quindi condurre anche a pesi finali negativi, i quali potrebbero essere non accettabili in alcune applicazioni.

La (19) è la funzione di distanza che viene utilizzata per l'ottenimento delle stime di massima verosimiglianza dei modelli log-lineari (Darroch and Ratcliff, 1972) ed è stata adottata in ambito ISTAT per il calcolo dei pesi finali dell'indagine multiscopo sulle Famiglie (ISTAT ,1993). Essa porta a coefficenti di correzione che possono variare nell'intervallo (0 , ∞) e conduce quindi a pesi finali sempre positivi. Tuttavia, in alcuni casi non favorevoli, i pesi finali possono presentare valori estremamente grandi rispetto ai corrispondenti pesi base, risultando pertanto non accettabili in quanto la loro applicazione per l'ottenimento di stime riferite a varie sottopopolazioni in differenti domini di studio può condurre a valori non realistici delle stime stesse.

La funzione di distanza logaritmica troncata conduce a pesi finali compresi nell'intervallo (Ld_k , Ud_k); questa caratteristica importante permette in primo luogo di ottenere pesi finali sempre positivi ponendo $L \geq 0$. Tale funzione di distanza rappresenta, in effetti, una funzione di distanza di tipo generalizzato in quanto al variare dei parametri L ed U prescelti dall'utente è possibile approssimare le soluzioni ottenute in base alle altre funzioni di distanza. Scegliendo un valore di L negativo e molto grande in valore assoluto ed un valore di U molto grande (ad esempio, L=-1.000, U=1000), la soluzione trovata approssima quella data dalla funzione di distanza lineare; con un valore di L positivo e molto piccolo ed un valore di U molto grande (ad esempio L=0,0001, U=1000) si approssima la soluzione data dalla funzione di distanza logaritmica.

2. STIMATORE DI REGRESSIONE GENERALIZZATA

2.1. Premessa

Il presente capitolo è finalizzato ad illustrare le principali caratteristiche logiche ed algebriche dello stimatore di *regressione generalizzata*, che si fonda sull'utilizzazione di variabili ausiliarie per le quali si conoscono i totali riferiti alla popolazione oggetto d'indagine. Ai fini della costruzione dello stimatore, si adopera l'informazione ausiliaria disponibile ipotizzando un modello di regressione che lega le variabili ausiliarie, che costituiscono le variabili esplicative del modello, alle variabili d'interesse. Una delle caratteristiche più interessanti dello stimatore in oggetto, è quella di poter utilizzare modelli regressivi con caratteristiche differenti; come vedremo meglio in seguito, ciò significa qualificare il modello regressivo in termini dei tre elementi fondamentali di: *gruppo di riferimento del modello*, *livello del modello*, *tipo di modello* [5].

Si dice *gruppo di riferimento del modello* un sottoinsieme (o sottopopolazione) della popolazione oggetto d'indagine con riferimento al quale:

- sono noti i totali della popolazione di una o più variabili ausiliarie;
- viene costruito il modello di regressione sottostante lo stimatore.

I *gruppi* rappresentano, quindi, una partizione della popolazione di riferimento e per ciascuno di essi si definisce uno specifico modello di regressione.

Da quanto detto risulta chiaro che, nella definizione del modello lineare, si possono utilizzare differenti specificazioni dei gruppi di riferimento del modello. E' possibile, infatti, definire i gruppi sia sulla base della partizione più fine (ossia la partizione che contiene più gruppi) rispetto alla quale sono noti i totali delle variabili ausiliarie, che sulla base di aggregazioni definite a partire dalla

[5] I tre elementi fondamentali che caratterizzano il modello di regressione corrispondono ai concetti di *model group*, *model level* e *model type*, introdotti nell'articolo di Estevao, Hidiroglou e Särndal (1995 §).

partizione più fine. Un caso particolare si ha quando l'intera popolazione definisce l'unico gruppo di riferimento del modello.

Il concetto di *livello del modello* è relativo al tipo di unità utilizzata nella formulazione del modello. Se le unità sulle quali è definito il modello di regressione sono costituite dai singoli elementi della popolazione, il modello è definito al *livello di unità elementari*; in tal caso, le variabili di interesse e quelle ausiliarie si riferiscono ai singoli elementi della popolazione. Se invece, le unità su cui è definito il modello sono costituite da gruppi o *cluster* di singoli elementi della popolazione, il modello è definito al *livello di cluster*; le variabili di interesse e quelle ausiliarie, si riferiscono, quindi, a *cluster* di elementi della popolazione.

Per i disegni di campionamento casuale semplice e ad uno stadio in cui si selezionano direttamente le singole unità della popolazione, il modello deve essere necessariamente definito al livello di elemento; questo, ad esempio, è il caso delle indagini ISTAT sulle imprese in cui il modello è definito al livello di impresa e i totali noti, riferiti a ciascun gruppo di riferimento del modello, sono costituiti, in genere, dal numero di imprese e dal numero totale di addetti appartenenti a tali imprese. Per i disegni ad uno stadio in cui si estraggono cluster di unità della popolazione e per i disegni di campionamento a due o più stadi di selezione il modello può essere definito sia al livello di elemento che al livello di cluster di elementi. Nei modelli definiti al livello di elemento i totali noti devono riferirsi a gruppi di singoli elementi mentre nei modelli al livello di cluster di elementi i totali noti devono riferirsi a gruppi di cluster.

Per quanto riguarda il concetto di *tipo di modello*, esso viene essenzialmente definito dal numero e dal tipo di variabili ausiliarie specificate nel modello e permette di definire i principali stimatori utilizzati in pratica nelle indagini campionarie.

Per illustrare le caratteristiche degli stimatori di regressione generalizzata, introduciamo livelli crescenti di complessità: a tal fine, nel successivo *paragrafo 2.2* viene introdotto lo stimatore di regressione generalizzata con riferimento al caso più semplice di modello definito al livello di elemento e di un unico gruppo di riferimento del modello, costituito da tutta la popolazione. Successivamente, nel *paragrafo 2.3*, dopo aver approfondito il concetto di *livello del modello*, viene

introdotto il modello di regressione al livello di cluster. Nel *paragrafo 2.4* è sviluppato il tema del *gruppo di riferimento del modello*; infine, nel *paragrafo 2.5* viene svolta una trattazione più approfondita del concetto di *tipo di modello*.

2.2. Modello a livello di unità elementari

2.2.1. Simbologia e prima formulazione dello stimatore di regressione generalizzata

Sia U una popolazione finita di N unità elementari, che indichiamo come $U = \{1,..., k,..., N\}$, e sia *s* un campione casuale di n elementi, che indichiamo come $s = \{1,..., k,..., n\}$, estratto da U mediante un disegno di campionamento che genera l'*universo dei campioni S* (ossia l'insieme di tutti i possibili campioni estraibili mediante il disegno in parola) ed assegna al generico campione s la probabilità *p*(*s*) di essere estratto, dove $\sum_{s \in S} p(s) = 1$. Con riferimento alla generica unità $k \in U$, indichiamo quindi con: $\pi_k = \sum_{s \in S(k)} p(s)$, la probabilità di inclusione nel campione dell'unità, dove *S*(k) denota il sottoinsieme di *S* caratterizzato dai campioni contenenti l'unità in oggetto; y_k, il valore assunto dalla variabile di interesse y; $x_k = (x_{1k},..., x_{jk},..., x_{Jk})$, il valore assunto dal vettore $\boldsymbol{x} = (x_1,..., x_j,..., x_J)'$ di J variabili ausiliarie.

Si vuole stimare il totale Y della variabile y , dato dalla seguente espressione:

$$Y = \sum_{k \in U} y_k \qquad (1)$$

sulla base delle seguenti informazioni:

- per ciascun elemento del campione s si dispone delle J+1 osservazioni $(y_k, \mathbf{x}_k)$;
- risultano conosciuti i J valori del vettore $X = (X_1,...,X_j,....,X_J)'$ dei totali delle J variabili ausiliarie, in cui

$$X_j = \sum_{k \in U} x_{jk} \qquad (j=1,...,J). \qquad (2)$$

E' utile chiarire che le J variabili ausiliarie vengono individuate cercando, nell'insieme delle variabili per le quali sono noti i totali al livello di popolazione, le variabili maggiormente correlate con la variabile y di interesse; in tal modo, il vettore di variabili ausiliarie ***x*** fornisce informazioni sulla variabile y di cui si può tenere conto nella fase di costruzione dello stimatore.

Introduciamo, quindi, un modello di regressione lineare, che indichiamo con ξ, per spiegare la forma della nuvola di punti definita sugli N elementi della popolazione finita U

$$\{(y_k, x_{1k},..., x_{jk},..., x_{Jk}): k = 1,....., N\}. \qquad (3)$$

Il modello si basa sulle seguenti assunzioni:

i) i valori $y_1,..., y_k,..., y_N$ assunti dalla variabile y per le N unità della popolazione sono considerati come realizzazioni di N variabili casuali indipendenti;

ii) le variabili ausiliarie sono trattate come costanti note di tipo non stocastico;

iii) la relazione che lega la generica variabile casuale y_k con il vettore $\mathbf{x}_k$ (k=1,.....,N) è la seguente

$$y_k = \mathbf{x}_k'\beta + \varepsilon_k \ , \quad (k=1...,N) \qquad (4)$$

in cui

$$\boldsymbol{\beta} = \left(\beta_1,...,\beta_j,...,\beta_J\right)'$$

è il vettore dei J coefficienti di regressione incogniti ed ε_k è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello ξ sono definiti rispettivamente da

$$E_\xi(\varepsilon_k) = 0\,,\ \mathrm{Var}_\xi(\varepsilon_k) = c_k\sigma^2\,,\ \mathrm{Cov}_\xi(\varepsilon_k,\varepsilon_l) = 0 \quad \mathrm{per}\ \forall\, k \neq l\,; \tag{5}$$

essendo $c_k\ (\mathrm{per}\, k \in U)$ delle costanti note.

iv) dalle precedenti relazioni (4) e (5) risultano, quindi, definiti anche i momenti, sotto il modello ξ, della generica variabile casuale y_k (k=1,...,N)

$$E_\xi(y_k) = \mathbf{x}_k'\boldsymbol{\beta}\ ,\ \mathrm{Var}_\xi(y_k) = c_k\sigma^2\ ,\ \mathrm{Cov}_\xi(y_k,y_l) = 0 \quad \mathrm{per}\ k \neq l\ . \tag{6}$$

Per quanto riguarda la varianza dei residui, ε_k, facciamo notare che nella formulazione adottata, riportata nella (5), è richiesta unicamente la conoscenza (o la stima) delle costanti c_k ma non quella del parametro σ^2, in quanto tale parametro si semplifica nella risoluzione del problema di regressione. Esempi di definizione delle costanti c_k sono: il caso di omoschedasticità dei residui in cui si pone c_k=1 (per $k \in$ s); oppure il caso in cui si dispone di un'unica variabile ausiliaria $\mathbf{x}_k = \mathbf{x}_{k1}$ e la variabilità di *y* tende ad aumentare all'aumentare dei valori di $\mathbf{x}_{k1}$, in tale situazione ha senso, quindi, porre $c_k = f(\mathbf{x}_{k1})$ (per $k \in$ s).

Ciò premesso, si supponga di aver effettuato un censimento di tutte le N unità della popolazione U e di disporre, quindi, di tutti i valori della nuvola di punti (3), si supponga, inoltre, che la nuvola di punti osservata si adatti piuttosto bene al modello ξ appena introdotto. E' possibile utilizzare, allora, la nuvola di punti

della popolazione per stimare, mediante il metodo dei minimi quadrati ponderati il vettore dei coefficienti di regressione β del modello ξ. Utilizzando la teoria standard della regressione generalizzata, si ha che il miglior stimatore lineare non distorto dei coefficienti β, sotto il modello ξ, è dato da

$$\mathbf{B} = \left(B_1,...,B_j,...,B_J\right)' = \left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \sum_{k\in U} \frac{\mathbf{x}_k y_k}{c_k}. \qquad (7)$$

Il vettore[6] dei coefficienti **B** è, ovviamente, una caratteristica incognita della popolazione. E' possibile, tuttavia, stimare **B**, mediante i dati rilevati dal campione s. La relazione (7) si presenta come il prodotto di totali della popolazione; ed una sua stima asintoticamente corretta può essere ottenuta, stimando correttamente ciascun totale mediante lo stimatore di Horvitz-Thompson. Siano, infatti

$$\mathbf{T}_1 = \sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \text{ e } \mathbf{T}_2 = \sum_{k\in U} \frac{\mathbf{x}_k y_k}{c_k} \qquad (8)$$

le due matrici che formano il secondo membro della (7), e siano rispettivamente

$$t_{1jj'} = \sum_{k\in U} \frac{x_{jk}\, x_{j'k}}{c_k} \quad \text{e} \quad t_{2j} = \sum_{k\in U} \frac{x_{jk}\, y_k}{c_k} \quad , \ (j \text{ e } j' = 1,...,J) \quad (9)$$

i generici elementi che formano tali matrici. Una stima corretta delle matrici (8) è data pertanto da

---

[6] La stima (7) è derivata mediante il metodo dei minimi quadrati, che, come è noto porta a definire il migliore stimatore lineare corretto.

$$\tilde{\mathbf{T}}_1 = \sum_{k\in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \frac{1}{\pi_k} \text{ e } \tilde{\mathbf{T}}_2 = \sum_{k\in s} \frac{\mathbf{x}_k y_k}{c_k} \frac{1}{\pi_k} \qquad (10)$$

i cui generici elementi sono espressi come stime dei totali (9) rispettivamente da

$$\tilde{t}_{1jj'} = \sum_{k\in s} \frac{x_{jk}\, x_{j'k}}{\pi_k\, c_k} \quad \text{e} \quad \tilde{t}_{2j} = \sum_{k\in s} \frac{x_{jk}\, y_k}{\pi_k c_k} \quad , \ (j \text{ e } j' = 1,\ldots, J) \qquad (11)$$

In sintesi, quindi, una stima asintoticamente corretta[7] della (7) è data da:

$$\tilde{\mathbf{B}} = \left(\tilde{B}_1,\ldots,\tilde{B}_j,\ldots,\tilde{B}_J\right)' = \tilde{\mathbf{T}}_1^{-1}\tilde{\mathbf{T}}_2 = \left(\sum_{k\in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k}\right)^{-1} \sum_{k\in s} \frac{\mathbf{x}_k y_k}{\pi_k c_k} . \qquad (12)$$

Per poter calcolare $\tilde{\mathbf{B}}$ mediante la (12), tutte le quantità indicate nella formula stessa devono essere note. Devono essere conosciuti, in particolare, i valori da assegnare alle quantità $\{c_k\}$ $(\text{per } k \in s)$.

Avendo attribuito un valore alle costanti $\{c_k\}$ $(\text{per } k \in s)$, data la nuvola di punti osservata per il campione s

$$\left\{\left(y_k, x_{1k},\ldots, x_{jk},\ldots, x_{Jk}\right): k = 1,\ldots..,s\right\} \qquad (13)$$

l'adattamento del modello ξ mediante i dati campionari rilevati porta a calcolare la stima dei coefficienti di regressione, $\tilde{\mathbf{B}}$, del modello attraverso la relazione (12). Sulla base di $\tilde{\mathbf{B}}$ e' possibile, quindi, calcolare:

---

[7] La stima è solo *asintoticamente* corretta in quanto il valore atteso dell'inversa di una matrice ad elementi casuali è diverso dall'inversa del valore atteso della matrice stessa. In formule ciò è espresso da $E(\tilde{T}_1) = T_1$, $E(\tilde{T}_1^{-1}) \neq \tilde{T}_1^{-1}$.

a) con riferimento alle N unità della popolazione, i valori interpolati $\tilde{y}_1,...,\tilde{y}_k,...,\tilde{y}_N$, relativi ai corrispondenti valori $y_1,...,y_k,...,y_N$, mediante la relazione

$$\tilde{y}_k = \mathbf{x}_k' \tilde{\mathbf{B}} = \sum_{j=1}^{J} \tilde{B}_j x_{jk} \quad \text{per } (k = 1,....., N) \qquad (14)$$

b) con riferimento alle n unità del campione i residui

$$e_k = y_k - \tilde{y}_k = y_k - \mathbf{x}_k' \tilde{\mathbf{B}} \quad \text{per } (k = 1,...., n). \qquad (15)$$

Ciò premesso, il totale di interesse Y può, quindi, essere riscritto mediante la seguente espressione

$$Y = \sum_{k \in U} y_k = \sum_{k \in U} \tilde{y}_k + \sum_{k \in U} (y_k - \tilde{y}_k) = \sum_{k \in U} \tilde{y}_k + \sum_{k \in U} e_k \qquad (16)$$

Le quantità appena introdotte sono alla base dello stimatore di regressione generalizzata, dall'analisi della (16) si osserva, infatti, che l'ultima relazione dopo il segno di uguaglianza è costituita dalla somma di due totali: il primo è una quantità nota, in quanto il valore di $\tilde{y}_k$ può essere definito per tutte le unità della popolazione; il secondo, invece, rappresenta una quantità incognita; non è possibile, infatti, calcolare i residui e_k per tutte le unità della popolazione ma solo per quelle appartenenti al campione osservato. Sostituendo, quindi, nella (16) lo stimatore corretto di Horvitz-Thompson di tale totale incognito, si ottiene lo stimatore di regressione generalizzata del totale Y, dato dalla seguente espressione

$$\tilde{Y}_{REG} = \sum_{k \in U} \tilde{y}_k + \sum_{k \in s} \frac{e_k}{\pi_k} \; . \qquad (17)$$

Dalla (17) risulta che lo stimatore di regressione generalizzata può essere espresso mediante la somma di due totali. Il primo è la somma degli N valori della popolazione stimati in base alla relazione (14) e contiene l'informazione ausiliaria disponibile al livello dei totali noti, $\mathrm{X}=(\mathrm{X}_1,...,\mathrm{X}_\mathrm{j},....,\mathrm{X}_\mathrm{J})'$, della popolazione. Il totale $\sum_{\mathrm{k}\in\mathrm{U}}\tilde{\mathrm{y}}_\mathrm{k}$ può essere, infatti, riscritto mediante il seguente passaggio:

$$\sum_{\mathrm{k}\in\mathrm{U}}\tilde{\mathrm{y}}_\mathrm{k}=\sum_{\mathrm{k}\in\mathrm{U}}\mathbf{x}_\mathrm{k}^{'}\tilde{\mathbf{B}}=\left(\sum_{\mathrm{k}\in\mathrm{U}}\mathbf{x}_\mathrm{k}\right)^{'}\tilde{\mathbf{B}}=\mathbf{X}^{'}\tilde{\mathbf{B}} \qquad (18)$$

Il secondo totale contenuto nella (17), è un termine di aggiustamento calcolato come somma pesata, con i pesi diretti π_k^{-1} , degli n dei residui campionari e_k e contiene l'informazione ausiliaria disponibile al livello delle singole unità campionarie, tale totale può essere, infatti, riscritto mediante il seguente passaggio

$$\sum_{\mathrm{k}\in\mathrm{s}}\tilde{\mathrm{e}}_\mathrm{k}=\sum_{\mathrm{k}\in\mathrm{s}}\frac{\left(\mathrm{y}_\mathrm{k}-\mathrm{x}_\mathrm{k}^{'}\tilde{\mathrm{B}}\right)}{\pi_\mathrm{k}}= \qquad (19)$$

$$=\sum_{\mathrm{k}\in\mathrm{s}}\left(\frac{\mathrm{y}_\mathrm{k}}{\pi_\mathrm{k}}\right)-\sum_{\mathrm{k}\in\mathrm{s}}\left(\frac{\mathbf{x}_\mathrm{k}}{\pi_\mathrm{k}}\right)^{'}\tilde{\mathbf{B}}=\tilde{\mathrm{Y}}-\tilde{\mathbf{X}}^{'}\tilde{\mathbf{B}},$$

in cui $\tilde{\mathrm{Y}}$ e $\tilde{\mathbf{X}}$ indicano le stime di Horvitz-Thompson dei corrispondenti totali noti, ottenute rispettivamente come

$$\tilde{\mathrm{Y}}=\sum_{\mathrm{k}\in\mathrm{s}}\left(\frac{\mathrm{y}_\mathrm{k}}{\pi_\mathrm{k}}\right) \quad , \quad \tilde{\mathbf{X}}=\sum_{\mathrm{k}\in\mathrm{s}}\left(\frac{\mathbf{x}_\mathrm{k}}{\pi_\mathrm{k}}\right)$$

Inserendo le precedenti formule (18) e (19) nella (17) è possibile, infine, esprimere lo stimatore di regressione generalizzato secondo l'espressione più usuale

$$\tilde{Y}_{REG} = \tilde{Y} + (\mathbf{X} - \tilde{\mathbf{X}})' \tilde{\mathbf{B}} \qquad (20)$$

dalla quale risulta che tale stimatore è ottenuto come somma dello stimatore di Horvitz-Thompson del totale Y più un termine di aggiustamento regressivo che dipende dalle differenze tra totali noti e corrispondenti stime campionarie di Horvitz-Thompson, ponderate con i rispettivi coefficienti di regressione stimati. Per calcolare lo stimatore di regressione generalizzata in base all'espressione (20) è necessario conoscere i totali delle variabili ausiliarie $X = (X_1,...,X_j,....,X_J)'$ ed i valori della variabile di interesse e delle variabili ausiliarie $\{(y_k, x_{1k},...,x_{jk},...,x_{Jk}) : k = 1,.....,s\}$ per le n unità campionate. Pertanto, non è necessario conoscere i valori delle variabili ausiliarie per le unità della popolazione non campionate[8].

Una fondamentale proprietà degli stimatori in parola è che la stima di regressione generalizzata dei totali delle variabili ausiliarie coincide con i valori conosciuti degli stessi. Infatti, sostituendo nella (20) le variabili ausiliarie $\mathbf{x}_k'$ alle variabili d'interesse y_k, si ha

$$\tilde{\mathbf{X}}'_{REG} = \tilde{\mathbf{X}}' + (\mathbf{X} - \tilde{\mathbf{X}})' \tilde{\mathbf{B}},$$

in cui introducendo l'espressione esplicita di $\tilde{\mathbf{B}}$ data dalla (12) si ottiene:

---

8 Questo è il caso delle indagini in cui sono disponibili unicamente i totali delle variabili ausiliarie mentre i singoli valori vengono rilevati con l'indagine campionaria. In tale caso, ovviamente, le variabili mediante le quali vengono costruiti i totali noti e le variabili rilevate sulle singole unità devono avere la medesima definizione concettuale ed essere riferite allo stesso periodo temporale. E' chiaro che l'allontanamento da tale condizione introduce nella stima fattori distorsivi.

$$\tilde{\mathbf{X}}'_{\mathrm{REG}} = \tilde{\mathbf{X}}' + (\mathbf{X} - \tilde{\mathbf{X}})' \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k} = \qquad (21)$$

$$= \tilde{\mathbf{X}}' + (\mathbf{X} - \tilde{\mathbf{X}})' = \mathbf{X}'.$$

2.2.2. Espressioni alternative dello stimatore di regressione generalizzata

Espressione in termini dei pesi

Un espressione alternativa dello stimatore di regressione generalizzata è data da

$$\tilde{Y}_{\mathrm{REG}} = \sum_{k \in s} y_k d_k \gamma_k = \sum_{k \in s} y_k w_k \ , \qquad (22)$$

in cui si è denotato con:

$$w_k = d_k \gamma_k ,$$ il *peso finale*,

$$d_k = \frac{1}{\pi_k} ,$$ il *peso diretto* anche detto *peso base*,

$$\gamma_k = 1 + (\mathbf{X} - \tilde{\mathbf{X}})' \left(\sum_{k \in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} \ ,$$ il *fattore correttivo del peso base*. (23)

La (22) permette di esprimere lo stimatore di regressione generalizzata come una somma ponderata, con i pesi $d_k \gamma_k$ (detti *pesi finali*), dei dati campionari $y_1,...,y_k,...,y_n$. Tale espressione è formalmente simile a quella dello stimatore di Horvitz - Thompson che, come è noto, è espresso come somma pesata, con i pesi base, dei dati campionari. Occorre, tuttavia, far notare che tra i due stimatori esiste una differenza sostanziale in quanto: i pesi diretti

dipendono unicamente dalle unità estratte nel campione e non dipendono dai valori delle variabili ausiliarie osservate nel campione; mentre i fattori γ_k e quindi i pesi finali, dipendono: i) dai totali noti delle variabili ausiliarie, ii) dai valori assunti dalle variabili ausiliarie nel campione estratto; iii) dalla variabilità della variabile oggetto di indagine. Un aspetto importante del vettore dei pesi finali pesi finali $\{w_k\}$, è quello che tali pesi possono assumere anche valori negativi; ciò può causare problemi logici in quanto il peso è strettamente connesso alla *rappresentatività di un unità* indicando quante unità non campionate della popolazione sono rappresentate dall'unità inclusa nel campione; di conseguenza, è molto problematico attribuire una rappresentatività ad un'unità che presenta un peso negativo. Inoltre la presenza di pesi negativi può causare l'effetto di definire valori negativi alle stime di totali di variabili che assumono valori sempre positivi o nulli.

Una interessante proprietà del vettore dei pesi finali pesi finali $\{w_k\}$ $(\text{per } k \in s)$ è che tale vettore rende minima la funzione di distanza *euclidea* tra l'insieme dei pesi diretti e quello dei pesi finali. Infatti, come viene esplicitato nel *paragrafo 1* è possibile esprimere lo stimatore di regressione generalizzato come uno stimatore della classe degli stimatori di *ponderazione vincolata* in quanto fattori γ_k possono essere ottenuti come soluzione del seguente problema di minimo vincolato:

$$\min\{G(d_k, w_k)\} = \min\left\{\frac{(d_k - d_k)^2}{d_k\, q_k}\right\}$$

in cui i vincoli sono dati da:

$$\sum_{s \in k} d_k \gamma_k\, x_k = X\,.$$

Da tale sistema si ottiene

$$\gamma_k = 1 + \mathbf{x}_k' \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\mathbf{X} - \tilde{\mathbf{X}} \right)$$

Espressione utile per il calcolo della varianza

Le espressioni alternative (17), (20) e (22) dello stimatore di regressione generalizzata sopra introdotte permettono di illustrare differenti aspetti e caratteristiche dello stimatore stesso. Tuttavia è importante aggiungere ad esse un'ultima espressione che sarà utile per derivare agevolmente la formula della varianza dello stimatore. A tal fine, si introducono le seguenti quantità

$$y_k^* = \mathbf{x}_k' \mathbf{B} \quad , \quad e_k^* = y_k - y_k^* \ \text{per} \ \ (k = 1, ..., N) \ , \qquad (24)$$

in cui **B** è calcolato mediante la (7); le quantità y_k^* rappresentano i valori teorici di *y* assunti dalle unità della popolazione, ottenuti interpolando nuvola di punti (3) costituita dagli N elementi della popolazione finita U attraverso la retta dei minimi quadrati ponderati e gli e_k^* rappresentano i residui calcolati dalla retta dei minimi quadrati in parola. Utilizzando le precedenti quantità è possibile modificare l'espressione (22) dello stimatore di regressione generalizzata, nel seguente modo

$$\tilde{Y}_{REG} = \sum_{k \in s} \frac{\gamma_k (y_k^* + e_k^*)}{\pi_k} = \sum_{k \in s} \frac{\gamma_k \, y_k^*}{\pi_k} + \sum_{k \in s} \frac{\gamma_k \, e_k^*}{\pi_k} \qquad (25)$$

Sulla base delle relazioni (24) e (21) è possibile riscrivere il primo addendo a secondo membro della (25) come

$$\sum_{k \in s} \frac{\gamma_k y_k^*}{\pi_k} = \sum_{k \in s} \left(\frac{\gamma_k \mathbf{x}_k'}{\pi_k} \right) \mathbf{B} = \mathbf{X}' \mathbf{B} = \sum_{k \in U} \mathbf{x}_k' \mathbf{B} = \sum_{k \in U} y_k^* \ . \qquad (26)$$

Sostituendo la (26) nella (25) si ottiene, infine, l'espressione cercata, data da

$$\tilde{Y}_{REG} = \sum_{k \in U} y_k^* + \sum_{k \in s} \frac{\gamma_k y_k^*}{\pi_k}. \qquad (27)$$

2.2.3. Alcune considerazioni sul ruolo del modello

E' importante svolgere alcune considerazioni circa il ruolo del modello ξ nella costruzione dello stimatore di regressione generalizzata.. Non è possibile, tuttavia, dare una dimostrazione formale delle proprietà enunciate poiché non è stata introdotta, ancora, la formula della varianza dello stimatore.

Lo stimatore è distorto ma è asintoticamente corretto, si ha quindi, che per campioni sufficientemente grandi (come quelli che caratterizzano le indagini effettuate dall'Istituto Nazionale di Statistica) si può assumere che lo stimatore sia corretto, inoltre lo stimatore è consistente nel senso che la sua varianza tende ad annullarsi al crescere della dimensione campionaria.

Le proprietà appena introdotte permettono di meglio comprendere la funzione giocata dal modello ξ, infatti esso ha essenzialmente la finalità di descrivere la nuvola di punti della popolazione finita. Si suppone, infatti, che il modello ξ costituisca una delle possibili *spiegazioni* della forma della nuvola di punti; non viene mai fatta, tuttavia, l'ipotesi che la popolazione sia stata realmente generata sulla base del modello in questione. L'introduzione del modello è, quindi, necessaria unicamente per definire una appropriata espressione di $\tilde{\mathbf{B}}$ da inserire nella formula dello stimatore di regressione.

L'efficienza dello stimatore in parola, in confronto a quella dello stimatore di Horvitz -Thompson, è funzione inversa dei residui e_k^*, e quindi dipende dal grado di adattamento della nuvola di punti della popolazione alla retta di regressione.

La proprietà di consistenza sotto il disegno delle stime ottenute mediante lo stimatore di regressione generalizzata e la validità della formula della varianza non dipendono, tuttavia, dal fatto se il modello sia valido oppure no. Da ciò deriva, in particolare, che l'inferenza prodotta è *assistita* dall'introduzione di un modello ma non è *dipendente* da esso in quanto tale (nella letteratura in lingua anglosassone, con riferimento ai due concetti appena introdotti, si usano rispettivamente i termini: *model assisted* e *model based*).

2.3. Livello del modello

2.3.1. Introduzione al problema

Nel precedente paragrafo abbiamo trattato il caso in cui il modello lineare alla base dello stimatore di regressione generalizzata è definito al *livello di unità elementare*, ovvero il caso in cui le variabili d'interesse e quelle ausiliarie si riferiscono ai singoli elementi della popolazione oggetto d'indagine ed i totali noti si ottengono come somma delle variabili ausiliarie sugli elementi della popolazione; tale tipo di modello è, ovviamente, l'unico che può essere utilizzato per i disegni ad uno stadio in cui si selezionano direttamente le singole unità della popolazione. Per i disegni ad uno stadio a *grappoli* - in cui si selezionano *grappoli* (o cluster) di singoli elementi, per i disegni a più stadi, è possibile definire sia modelli al *livello di elemento* che *modelli a livello di grappolo*. In tal caso le variabili d'interesse e quelle ausiliarie si riferiscono a grappoli di elementi ed i totali si ottengono come somma, sulla popolazione dei grappoli, delle variabili ausiliarie relative ai grappoli. Per meglio illustrare tale aspetto consideriamo gli esempi di seguito riportati.

Esempio 1

La popolazione oggetto di indagine è costituita dagli individui; si effettua un piano di campionamento a due stadi in cui si selezionano al primo stadio i

comuni e al secondo stadio le famiglie e si osservano le variabili oggetto di indagine su tutti gli individui appartenenti alle famiglie campione. Le variabili ausiliarie, riferite agli individui, sono il *sesso* e l'*età*; i totali noti sono definiti dalla distribuzione della popolazione per sesso ed età. Nella situazione appena descritta gli individui sono le *unità elementari* e le famiglie costituiscono *grappoli di unità.*

Esempio 2

La popolazione oggetto di indagine è costituita dalle unità locali; si adotta un piano di campionamento ad uno stadio a grappoli in cui si selezionano le imprese e si osservano le variabili oggetto di indagine su tutte le unità locali appartenenti alle imprese campione. Le variabili ausiliarie, sono i dati fiscali dell'impresa e non è possibile disporre di tali dati a livello di singola unità locale. I totali noti sono costituiti dai totali dei dati fiscali sulla popolazione delle imprese. Nella situazione appena descritta le unità locali sono le *unità elementari* e le imprese sono *grappoli di unità.*

Nel campionamento a grappolo è possibile evidenziare due situazioni distinte relativamente alla disponibilità delle informazioni ausiliarie:

a) le informazioni ausiliarie sono disponibili a livello di elemento (vedi esempio 1);

b) le informazioni ausiliarie sono disponibili solamente al livello di grappolo mentre non sono note tali informazioni per ciascun elemento appartenente al grappolo (vedi esempio 2).

Nella situazione a) è possibile definire sia un modello a *livello di elemento* sia un modello a *livello di grappolo* aggregando le informazioni ausiliarie delle unità elementari del grappolo; viceversa nella situazione b) è possibile definire solo un modello a livello di grappolo.

Nel caso in cui si adottino disegni a più stadi di campionamento (come nell'esempio 1), a seconda della disponibilità dell'informazione ausiliaria, è

possibile definire il livello del modello in modo differente: ad esempio è possibile individuare:

(i) un modello a livello di elemento;

(ii) un modello a livello di unità primaria;

(iii) un modello a livello di grappoli di unità elementari selezionati all'ultimo stadio di campionamento;

(iv) un modello a più livelli di riferimento. Ad esempio a livello di elemento e a livello di unità primaria.

Per illustrare il modello a più livelli, riprendiamo l'esempio 1, ed ipotizziamo di conoscere la zona altimetrica per comune. In tale situazione è possibile definire un modello a più livelli utilizzando l'informazione ausiliaria riferita agli individui e quella relativa ai comuni.

Nel seguente paragrafo illustreremo come viene definito il modello a livello di grappolo, mentre nel *paragrafo 2.3.3* descriveremo il caso del modello a livello di unità primaria.

2.3.2. Campionamento a grappoli

Consideriamo una popolazione U di N elementi ripartita in N_I grappoli ed indichiamo con: i l'indice di grappolo; $U_I = \{1,...,i,...,N_I\}$ la popolazione dei grappoli; N_{Ii} il numero delle unità elementari del grappolo i-esimo; k l'indice di unità elementare (k=1,..., N_{Ii}) . Supponiamo di avere estratto da U_I un campione casuale mediante il seguente schema:

(i) si seleziona un campione $s_I = \{1,...,i,...,n_I\}$ di n_I grappoli mediante il disegno di campionamento che genera l'universo dei campioni S_I e assegna al generico campione s_I la probabilità $p_I(s_I)$ di essere estratto (dove $\sum_{s_I \in S_I} p(s_I) = 1$);

(ii) di conseguenza, indicando con $S_I(i)$ il sottoinsieme di S_I formato dai campioni contenenti il grappolo i-esimo la probabilità d'inclusione di tale grappolo è data da $\pi_{Ii} = \sum_{s_I \in S(i)} p_I(s_I)$;

(iii) tutte le unità elementari dei grappoli selezionati vengono incluse nel campione; tale circostanza determina il fatto che la probabilità d'inclusione delle unità elementari coincide con quella dei grappoli di appartenenza; pertanto, la probabilità di inclusione π_{Iik} dell'unità elementare k appartenente al grappolo i è data da $\pi_{Iik} = \pi_{Ii}$ (per k=1,...,N_{Ii}, e per i=1,..., N_I).

Facendo riferimento al k-esimo elemento del grappolo i (per k=1,...,N_{Ii}, e per i=1,..., N_I) indichiamo quindi con y_{Iik} il valore della variabile d'interesse *y* e con $\mathbf{x}_{Iik}$ il valore assunto dal vettore di J variabili ausiliarie; considerando, quindi, il grappolo[9] nel suo complesso si ha:

$$y_{Ii} = \sum_{k=1}^{N_{Ii}} y_{Iik} \quad \text{e} \quad \mathbf{x}_{Ii} = \sum_{k=1}^{N_{Ii}} \mathbf{x}_{Iik} .$$

Modello a livello di grappolo

Per stimare, il totale Y della variabile d'interesse *y* definito da

$$Y = \sum_{i=1}^{N_I} \sum_{k=1}^{N_{Ii}} y_{Iik} = \sum_{i=1}^{N_I} y_{Ii} , \qquad (28)$$

utilizziamo il seguente modello ξ a livello di grappolo

$$y_{Ii} = \mathbf{x}_{Ii}^{'} \boldsymbol{\beta}_I + \varepsilon_{Ii} \qquad \text{per } (i=1,..., N_I) \qquad (29)$$

---

9 E' chiaro che nella situazione b), illustrata nel precedente *paragrafo 2.3.1*, è possibile definire unicamente il valore delle variabili ausiliarie a livello di grappolo $\mathbf{x}_{Ii}$ e non il valore $\mathbf{x}_{Iik}$ per ciascuno degli elementi del grappolo.

dove $\boldsymbol{\beta}_I = \left(\beta_{I1},...,\beta_{Ij},...,\beta_{IJ}\right)'$ è il vettore dei J coefficienti di regressione incogniti ed ε_{Ii} è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello sono definiti rispettivamente da

$$E_\xi(\varepsilon_{Ii}) = 0\,,\ Var_\xi(\varepsilon_{Ii}) = c_{Ii}\sigma^2\,,\ Cov_\xi(\varepsilon_{Ii},\varepsilon_{Ii'}) = 0 \quad \text{per}\ \forall\, i \neq i'\,; \tag{30}$$

essendo c_{Ii} (per i=1,..., N_I) delle costanti note.

Sotto il modello appena introdotto lo stimatore di regressione generalizzata è dato da

$$\begin{aligned}\tilde{Y}_{REG} &= \sum_{i=1}^{n_I} y_{Ii} d_{Ii}\ \gamma_{Ii} \\ &= \sum_{i=1}^{n_I} y_{Ii} w_{Ii} \quad , \\ &= \sum_{i=1}^{n_I} w_{Ii} \sum_{k=1}^{N_{Ii}} y_{Iik}\end{aligned} \tag{31}$$

in cui si è denotato con:

$w_{Ii} = d_{Ii}\ \gamma_{Ii}\,,$ il *peso finale*,

$d_{Ii} = \dfrac{1}{\pi_{Ii}}\,,$ il *peso diretto*,

$\gamma_{Ii} = 1 + (\mathbf{X}_I - \tilde{\mathbf{X}}_I)' \left(\sum_{i=1}^{n_I} \dfrac{d_{Ii}\mathbf{x}_{Ii}\mathbf{x}_{Ii}'}{c_{Ii}} \right)^{-1} \dfrac{\mathbf{x}_{Ii}}{c_{Ii}}$ il *fattore correttivo del peso base*,

essendo

$$\mathbf{X}_I = \sum_{i=1}^{N_I} \mathbf{x}_{Ii} \quad , \quad \tilde{\mathbf{X}}_I = \sum_{i=1}^{n_I} \mathbf{x}_{Ii}\, d_{Ii} \,.$$

Definire un modello a livello di grappolo comporta, quindi, il fatto di assegnare il peso finale del grappolo anche a tutte le unità elementari ad esso appartenenti.

Modello a livello di unità elementare

Nel caso in cui sia noto il valore del vettore delle variabili ausiliarie $\mathbf{x}_{Iik}$ per ciascun elemento di ogni grappolo - come nel caso dell'esempio 1 del par. 3.1- è possibile definire in alternativa a quanto appena illustrato un modello al *livello di elemento*, analogo a quello definito dalle relazioni (4) e (5) del *paragrafo 2.2.1*

$$y_{Iik} = \mathbf{x}'_{Iik}\boldsymbol{\beta}_I + \varepsilon_{Iik} \qquad (32)$$

dove $\boldsymbol{\beta}_I$ è il vettore dei coefficienti di regressione incogniti ed ε_{Iik} è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello sono definiti rispettivamente da

$$E_\xi(\varepsilon_{Iik}) = 0\,, \quad Var_\xi(\varepsilon_{Iik}) = c_{Iik}\sigma^2\,, \quad Cov_\xi(\varepsilon_{Iik}, \varepsilon_{Ii'k'}) = 0 \quad \text{per} \; \forall\, ik \neq i'k'\,; \qquad (33)$$

essendo c_{Iik} delle costanti note. In base al modello appena introdotto è quindi possibile derivare lo stimatore di regressione generalizzato come illustrato nel *paragrafo 2*

$$\tilde{Y}_{REG} = \sum_{i=1}^{n_I} \sum_{k=1}^{N_{Ii}} y_{Iik} d_{Ii}\, \gamma_{Iik} = \sum_{i=1}^{n_I} \sum_{k=1}^{N_{Ii}} y_{Iik} w_{Iik} \quad , \qquad (34)$$

in cui si è denotato con

$$w_{Iik} = d_{Ii}\,\gamma_{Iik},$$ il *peso finale*,

$$d_{Ii} = \frac{1}{\pi_{Ii}},$$ il *peso diretto*,

$$\gamma_{Iik} = 1 + (\mathbf{X} - \tilde{\mathbf{X}})' \left(\sum_{i=1}^{n_I} \sum_{k=1}^{N_{Ii}} \frac{d_{Ii}\mathbf{x}_{Iik}\mathbf{x}'_{Iik}}{c_{Iik}} \right)^{-1} \frac{\mathbf{x}_{Iik}}{c_{Iik}}$$ il *fattore correttivo del peso base*

dove

$$\mathbf{X} = \sum_{i=1}^{N_I} \sum_{k=1}^{N_{Ii}} \mathbf{x}_{Iik} \quad , \quad \tilde{\mathbf{X}} = \sum_{i=1}^{n_I} \sum_{k=1}^{N_{Ii}} \mathbf{x}_{Ii}\, d_{Ii}\,.$$

La definizione di un modello a livello di unità elementare comporta, quindi, il fatto che a ciascuna unità di un grappolo venga attribuito un peso finale differente[10].

Scelta del livello del modello

Un campionamento a grappoli consente di stimare parametri riferiti sia alla popolazione delle unità elementari che alla popolazione dei grappoli. Per illustrare tale aspetto consideriamo, ad esempio, un indagine sulle famiglie in cui le famiglie costituiscono i grappoli e gli individui le unità elementari e supponiamo di avere rilevato per ciascuna famiglia una variabile dicotomica y_{Ii} che assume valore 1 nel caso che la famiglia abbia tre componenti e valore 0 altrimenti; supponiamo, inoltre, di avere rilevato per ogni individuo una variabile dicotomica y_{Iik} che assume valore 1 nel caso in cui esso viva in una famiglia di tre componenti e valore 0 altrimenti; utilizzando la variabile y_{Ii} è possibile ottenere una stima del numero di famiglie con tre componenti, mentre utilizzando le variabili y_{Iik} si può calcolare la stima del numero di persone che vivono in

10 Questo non è vero nel caso in cui tutte le unità del grappolo presentino lo stesso valore delle variabili ausiliarie.

famiglie di tre componenti; è ovvio che questa ultima stima divisa per tre fornisce, nuovamente, una stima del numero di famiglie di tre componenti. L'esempio appena introdotto, mostra un caso molto frequente nelle indagini ISTAT sulle famiglie in cui è possibile ottenere una stima di uno stesso parametro oggetto di indagine (ad esempio, il numero di famiglie di tre componenti) sia utilizzando le informazioni relative agli individui sia quelle relative alle famiglie; nasce da qui l'esigenza di *coerenza* tra l'insieme delle stime riferite alle unità elementari e quelle riferite ai grappoli. Tale coerenza, implica quindi l'uguaglianza delle stime relative allo stesso parametro incognito di popolazione e si ottiene attribuendo il medesimo peso finale al grappolo ed a tutte le unità elementari ad esso appartenenti.

Da quanto appena illustrato si desume che la scelta del livello del modello è strettamente dipendente dagli obiettivi dell'indagine. Per una generica indagine che utilizza il campionamento a grappolo, possiamo evidenziare i tre seguenti tipi di obiettivo:

1. stimare unicamente parametri riferiti alla popolazione delle unità elementari;
2. stimare unicamente parametri riferiti alla popolazione dei grappoli;
3. stimare parametri riferiti sia alla popolazione delle unità elementari che a quella dei grappoli.

Nel primo caso, in cui si stimano parametri riferiti alle unità elementari, è possibile adottare sia un modello a livello di unità elementare che un modello a livello di grappolo; ha senso, pertanto, adottare il modello che garantisce la minimizzazione degli errori campionari.

Nel secondo caso, in cui si stimano parametri riferiti ai grappoli, è in genere auspicabile l'utilizzazione di un modello a livello di grappolo che costruisce direttamente un peso finale per il grappolo. Un modello a livello di elemento assegna un peso finale differente a tutte le unità elementari appartenenti al

grappolo; risulta quindi difficile attribuire un peso finale al grappolo differente dal suo peso diretto[11].

Nel terzo caso, in cui si stimano congiuntamente parametri riferiti ai grappoli e parametri relativi alle unità elementari, è in genere auspicabile l'utilizzazione di un modello a livello di grappolo che risolve i problemi di coerenza delle stime assegnando lo stesso peso finale al grappolo ed a tutte le unità elementari ad esso appartenenti. Nel caso in cui non si pongano problemi di coerenza, ossia nel caso in cui non sia possibile derivare le stime riferite ai grappoli da quelle calcolate per le unità elementari (o viceversa), si possono definire due modelli distinti: uno per i grappoli e l'altro per le unità elementari.

Sintetizziamo nella seguente tabella i criteri di scelta appena descritti.

Obiettivi dell'indagine	**Criterio di scelta**
Stime riferite alle unità elementari	modello a livello di grappolo o di unità elementare a seconda degli errori campionari delle stime
Stime riferite ai grappoli	modello a livello di grappolo
Stime riferite alle unità elementari e ai grappoli	modello a livello di grappolo

2.3.3. Disegni di campionamento a due o più stadi

Consideriamo una popolazione U di N elementi ripartita in N_I *Unità Primarie* (UP) di campionamento ed indichiamo con $U_I = \{1,...,i,...,N_I\}$ la popolazione delle unità primarie.

[11] Questo problema può essere risolto, anche, mediante il metodo noto nelle letteratura in lingua anglosassone con il termine *principal person method* che assegna al grappolo il peso finale calcolato per l'unità elementare più rappresentativa o più importante del grappolo Alexander (1987).

Ipotizziamo, inoltre, che la i-esima UP sia costituita da N_{Ii} *Unità Secondarie* (US). Le US possono essere unità elementari o alternativamente grappoli di unità elementari. Indicando

quindi con y_{Iik} il valore della variabile d'interesse *y* relativo alla US k-esima dell'UP i-esima (per k=1,...,N_{Ii}, e per i=1,..., N_I), il totale Y della variabile d'interesse è definito da

$$Y = \sum_{i=1}^{N_I} Y_{Ii} , \qquad (35)$$

essendo

$$Y_{Ii} = \sum_{k=1}^{N_{Ii}} y_{Iik}$$

il totale della variabile y riferito alla UP i (per $i = 1,..., N_I$).

Supponiamo, ora, di avere estratto da U_I un campione casuale mediante il seguente schema articolato in due stadi di campionamento:

(i) al primo stadio si seleziona un campione $s_I = \{1,..., i,..., n_I\}$ di n_I UP mediante un disegno di campionamento che genera l'universo dei campioni S_I e assegna al generico campione s_I la probabilità $p_I(s_I)$ di essere estratto (dove $\sum_{s_I \in S_I} p_I(s_I) = 1$); di conseguenza, indicando con $S_I(i)$ il sottoinsieme di S_I formato dai campioni contenenti la UP i-esima, la probabilità d'inclusione di tale UP è data da $\pi_{Ii} = \sum_{s_I \in S_I(i)} p_I(s_I)$;

(ii) al secondo stadio: dalla i-esima UP campione (per i=1,..., n_I), si seleziona un campione $s_{Ii} = \{1,..., k,..., n_{Ii}\}$ di n_{Ii} US mediante un meccanismo di selezione che genera l'universo dei campioni S_{Ii} e assegna al generico campione s_{Ii} la probabilità $p_{Ii}(s_{Ii})$ di essere estratto (dove $\sum_{s_{Ii} \in S_{Ii}} p(s_{Ii}) = 1$); pertanto, indicando con $S_{Ii}(k)$ il sottoinsieme di S_{Ii}

caratterizzato dai campioni contenenti la US k-esima (per $k = \{1,..., N_{Ii}\}$), si ha che per la US in parola la probabilità d'inclusione condizionata (alla selezione dell'US i-esima) è data da $\pi_{Iik|Ii} = \sum_{s_{Ii} \in S_{Ii}(k)} p_{Ii}(s_{Ii})$ (per k=1,...,N_{Ii}, e per i=1,..., N_I);

(iii) in conseguenza di quanto appena illustrato, la probabilità di inclusione finale dell'US k-esima appartenente all'UP i-esima è definita da $\pi_{Iik} = \pi_{Ii}\ \pi_{Iik|Ii}$ (per k=1,...,N_{Ii}, i=1,..., N_I).

Nel contesto campionario in parola, per stimare il totale Y della variabile d'interesse *y* è possibile definire sia modelli al livello di UP che modelli al livello di US. Per quanto riguarda i modelli al livello di US si può fare riferimento a quanto illustrato nel precedente paragrafo; se, infatti, le US sono unità elementari, un modello al livello di US corrisponde ad un modello a livello di unità elementari; se, invece, le US costituiscono grappoli di unità elementari è possibile adottare, in base agli obiettivi dell'indagine, sia un modello al livello di unità elementare che un modello a livello di grappolo.

Qui di seguito illustriamo il modo per ottenere le stime adottando il seguente modello a livello di UP:

$$Y_{Ii} = \mathbf{x}'_{Ii}\ \boldsymbol{\beta}_I + \varepsilon_{Ii} \qquad \text{per } (i=1,..., N_I) \qquad (36)$$

dove $\boldsymbol{\beta}_I$ è il vettore dei coefficienti di regressione incogniti e, con riferimento alla UP i-esima, $\mathbf{x}_{Ii}$ indica il vettore delle variabili ausiliarie (supposto noto) e ε_{Ii} è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello sono definiti rispettivamente da

$$E_\xi(\varepsilon_{Ii}) = 0\,,\ Var_\xi(\varepsilon_{Ii}) = c_{Ii}\sigma^2\,,\ Cov_\xi(\varepsilon_{Ii}, \varepsilon_{Ii'}) = 0 \quad \text{per} \forall\, i \neq i'\,; \qquad (37)$$

essendo c_{Ii} (per i=1,..., N_I) delle costanti note.

Utilizzando la stima corretta $\tilde{Y}_{Ii} = \sum_{i=1}^{n_{Ii}} \frac{y_{Iik}}{\pi_{Iik|Ii}}$ del totale Y_{Ii}, sulla base del modello (36) e (37) è possibile definire il seguente stimatore di regressione:

$$\begin{aligned} \tilde{Y}_{REG} &= \sum_{i=1}^{n_I} \tilde{Y}_{Ii} \frac{1}{\pi_{Ii}} \gamma_{Ii} \\ &= \sum_{i=1}^{n_I} \gamma_{Ii} \sum_{k=1}^{n_{Ii}} y_{Iik} d_{Iik} \quad , \\ &= \sum_{i=1}^{n_I} \sum_{k=1}^{n_{Ii}} y_{Iik} w_{Iik} \end{aligned} \qquad (38)$$

in cui si è denotato con:

$w_{Iik} = d_{Iik}\, \gamma_{Ii}$, il *peso finale*,

$d_{Iik} = \frac{1}{\pi_{Iik}}$, il *peso diretto*,

$\gamma_{Ii} = 1 + (\mathbf{X}_I - \tilde{\mathbf{X}}_I)' \left(\sum_{i=1}^{n_I} \frac{\mathbf{x}_{Ii} \mathbf{x}_{Ii}'}{\pi_{Ii} c_{Ii}} \right)^{-1} \frac{\mathbf{x}_{Ii}}{c_{Ii}}$, il *fattore correttivo del peso base*

essendo

$$\mathbf{X}_I = \sum_{i=1}^{N_I} \mathbf{x}_{Ii} \quad , \quad \tilde{\mathbf{X}}_I = \sum_{i=1}^{n_I} \frac{\mathbf{x}_{Ii}}{\pi_{Ii}} \ .$$

Dall'esame delle precedenti espressioni è possibile svolgere le seguenti considerazioni:

1. adottare un modello a livello di UP ha come conseguenza il fatto che tutte le US di una data UP presentino il medesimo valore del fattore correttivo del peso base;
2. dal punto precedente discende che il peso finale è uguale per tutte le US di una data UP solamente nel caso in cui: (a) si adotta un modello a livello di UP; (b) nel secondo stadio di campionamento le US sono selezionate con probabilità uguali;
3. nel caso in cui le variabili ausiliarie non sono conosciute a livello di UP ma è noto unicamente il valore $\mathbf{x}_{Iik}$ delle US campione, il fattore correttivo del peso base viene modificato sostituendo al posto del totale $\mathbf{x}_{Ii}$ una sua stima corretta data dall'espressione

$$\tilde{\mathbf{x}}_{Ii} = \sum_{i=1}^{n_{Ii}} \frac{\mathbf{x}_{Iik}}{\pi_{Iik|Ii}}.$$

Per quanto riguarda il problema della scelta del livello del modello, valgono le considerazioni illustrate nel *paragrafo 2.3.2*, secondo le quali il livello del modello deve essere scelto essenzialmente sulla base degli obiettivi dell'indagine. In questa sede facciamo notare che, qualora un indagine a due stadi abbia la finalità di produrre stime anche per la popolazione delle UP è necessario adottare una strategia che conduca ad assegnare pesi uguali a tutte le US di una data UP ossia, come già detto, una strategia campionaria in cui le US siano selezionate nel secondo stadio con probabilità uguale ed in cui si adotti un modello a livello di UP.

2.4. Gruppo di riferimento del modello

2.4.1. Modello a livello di unità elementare

Nel presente paragrafo riprendiamo l'importante concetto di *gruppo di riferimento del modello* che è stato già introdotto brevemente nel *paragrafo 2.2*. La trattazione verrà svolta dapprima per il caso di un modello *al livello di elemento*; l'estensione al caso di un modello a *livello di grappolo* sarà sviluppata nel successivo paragrafo.

Uno dei più importanti aspetti della caratterizzazione del modello di regressione, sottostante allo stimatore di regressione generalizzata, è legato alla possibilità di suddividere, sulla base di una o più variabili di classificazione, la popolazione U di N elementi in un certo numero, G, di *sottopopolazioni* (o *gruppi*), che indichiamo con i simboli $U_{(1)},..., U_{(g)},...,U_{(G)}$, contenenti rispettivamente $N_{(1)},..., N_{(g)},....N_{(G)}$ elementi della popolazione. Ciascuna delle *sottopopolazioni* così formate costituisce un *gruppo di riferimento del modello* se sono rispettate le seguenti condizioni:

- l'insieme dei *gruppi* è una *partizione completa* della popolazione U, ciò significa, in particolare, che l'intersezione di due gruppi differenti è sempre uguale all'insieme vuoto e che l'unione dei G gruppi coincide con la popolazione U. In simboli si ha quindi

$$U = \bigcup_{g=1}^{G} U_{(g)} \quad \text{e} \quad \varnothing = U_{(g)} \bigcap U_{(g')} \quad (\text{per } g \neq g' = 1,.....,G)$$

da cui deriva

$$\sum_{g=1}^{G} N_{(g)} = N;$$

– sono conosciuti i totali $\mathbf{X}_{(g)} = (X_{(g)1},...,X_{(g)J})'$ delle variabili ausiliarie per ciascun gruppo g essendo

$$\sum_{k=1}^{N_{(g)}} \mathbf{x}_k = \mathbf{X}_{(g)};$$

– il campione $s_{(g)}$ del gruppo g definito come $s_{(g)} = s \cap U_{(g)}$, deve essere costituito da un numero $n_{(g)}$ di unità elementari sempre maggiore del numero J di totali noti.

Valendo le precedenti condizioni è possibile definire un modello separato per le unità di ciascun gruppo, espresso come

$$y_k = \mathbf{x}_k' \boldsymbol{\beta}_{(g)} + \varepsilon_k \qquad \text{per } k \in U_{(g)} \qquad (39)$$

dove $\boldsymbol{\beta}_{(g)}$ è il vettore dei coefficienti di regressione incogniti ed ε_k è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello sono definiti rispettivamente da

$$E_\xi(\varepsilon_k) = 0,\ Var_\xi(\varepsilon_k) = c_k\sigma^2,\ Cov_\xi(\varepsilon_k, \varepsilon_{k'}) = 0 \ \text{per} \ \forall\, k \neq k'; \qquad (40)$$

essendo c_k delle costanti note. In base al modello appena introdotto è possibile, quindi, derivare lo stimatore di regressione generalizzata come illustrato nel *paragrafo 2.2*.

$$\tilde{Y}_{REG} = \sum_{g=1}^{G} \sum_{k=1}^{n_{(g)}} y_k d_k \gamma_k = \sum_{g=1}^{G} \sum_{k=1}^{n_{(g)}} y_k w_k, \qquad (41)$$

in cui si è indicato con:

$w_k = d_k \gamma_k$, il *peso finale*,

$d_k = \frac{1}{\pi_k}$, il *peso diretto*,

$\gamma_k = 1 + (\mathbf{X}_{(g)} - \tilde{\mathbf{X}}_{(g)})' \left(\sum_{k=1}^{n_{(g)}} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k}$, il *fattore correttivo del peso base*, di un unita campionaria appartenente al gruppo g.

$(k \in s_{(g)})$

essendo

$$\tilde{\mathbf{X}}_{(g)} = \sum_{k=1}^{n_{(g)}} \mathbf{x}_k d_k$$

la stima diretta del totale del vettore dei totali $\mathbf{X}_{(g)}$.

E' possibile dimostrare che definire un modello separato per ciascun gruppo g $(g = 1,...,G)$ è equivalente ad un modello lineare generale del tipo

$$y_k = \mathbf{z}_k' \boldsymbol{\beta} + \varepsilon_k \qquad \text{per } k \in U \qquad (42)$$

$$\mathbf{z}_k' = \left(\delta_{(1)k} \mathbf{x}_k', ..., \delta_{(g)k} \mathbf{x}_k', ..., \delta_{(G)k} \mathbf{x}_k' \right) \qquad (43)$$

$$E_\xi(\varepsilon_k) = 0,\ Var_\xi(\varepsilon_k) = c_k \sigma^2,\ Cov_\xi(\varepsilon_k, \varepsilon_{k'}) = 0 \ \text{per} \ \forall\, k \neq k'; \quad (44)$$

dove $\delta_{(g)k}$ (g = 1,...,G) è una variabile dicotomica che assume valore 1 se l'unità k-esima appartiene al gruppo g e valore 0 altrimenti e i vettori $\mathbf{z}_k'$ e $\boldsymbol{\beta}$ sono costituiti da A=JxG elementi. Il vettore $\boldsymbol{\delta}_k' = \left(\delta_{(1)k}, ..., \delta_{(g)k}, ..., \delta_{(G)k} \right)$ che

contiene le variabili indicatrici $\delta_{(g)k}$ appena introdotte ha J-1 termini pari a zero ed un singolo termine pari ad 1, che identifica il gruppo al quale il k-esimo l'elemento appartiene; è valida, pertanto, la seguente relazione

$$\sum_{k=1}^{N} \boldsymbol{\delta}_k^{'} = (N_{(1)},...,N_{(g)},...,N_{(G)})$$

essendo $(N_{(1)},...,N_{(g)},...,N_{(G)})$ il vettore contenente rispettivamente le numerosità della popolazione in ciascuno dei gruppi considerati.

Lo stimatore di regressione sotto il modello definito dalle (42) - (44) è dato da:

$$\tilde{Y}_{REG} = \sum_{k=1}^{n} y_k d_k \gamma_k = \sum_{k=1}^{n} y_k w_k \ , \qquad (45)$$

in cui si è denotato con:

$w_k = d_k \gamma_k$, il *peso finale*,

$\gamma_k = 1 + (\mathbf{Z} - \tilde{\mathbf{Z}})^{'} \left(\sum_{k=1}^{n} \frac{d_k \mathbf{z}_k \mathbf{z}_k^{'}}{c_k} \right)^{-1} \frac{\mathbf{z}_k}{c_k}$, il *fattore correttivo del peso base*

dove

$$\mathbf{Z} = \sum_{k=1}^{N} \mathbf{z}_k \quad , \quad \tilde{\mathbf{Z}} = \sum_{k=1}^{n} \mathbf{z}_k d_k \ .$$

essendo la matrice

$$\left(\sum_{k=1}^{n} \frac{d_k \mathbf{z}_k \mathbf{z}_k'}{c_k} \right)^{-1}$$

una matrice diagonale a blocchi in cui il generico blocco g (per g=1,...,G) è definito da

$$\mathbf{Q}_{(g)} = \left(\sum_{k=1}^{n(g)} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} .$$

Risulta facile dimostrare che lo stimatore di regressione espresso dalla (45) è uguale a quello definito dalla (41), infatti il correttore γ_k per una unità appartenente al gruppo g è definito da:

$$\gamma_k = 1 + \left[\mathbf{X}_{(1)}' - \tilde{\mathbf{X}}_{(1)}', ..., \mathbf{X}_{(g)}' - \tilde{\mathbf{X}}_{(g)}', ..., \mathbf{X}_{(G)}' - \tilde{\mathbf{X}}_{(G)}' \right] \begin{bmatrix} \mathbf{Q}_{(1)} & 0 & \cdots & & 0 \\ 0 & & & \cdots & \vdots \\ \vdots & 0 & \mathbf{Q}_{(g)} & 0 & \vdots \\ \vdots & & & & 0 \\ 0 & \cdots & \cdots & 0 & \mathbf{Q}_{(G)} \end{bmatrix} \frac{1}{c_k} \begin{bmatrix} 0 \\ \vdots \\ \mathbf{x}_k \\ \vdots \\ 0 \end{bmatrix}$$

$$= 1 + (\mathbf{X}_{(g)} - \tilde{\mathbf{X}}_{(g)})' \mathbf{Q}_{(g)} \frac{\mathbf{x}_k}{c_k}$$

$$= 1 + (\mathbf{X}_{(g)} - \tilde{\mathbf{X}}_{(g)})' \left(\sum_{k=1}^{n(g)} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} \quad (\text{per } k \in s_{(g)}).$$

I post-strati definiscono un importante caso di gruppi di riferimento del modello, poiché essi per definizione costituiscono delle sub-popolazioni non sovrapposte per le quali sono noti i totali di riferimento, altre sub-popolazioni spesso considerate nel formare i gruppi possono essere gli strati, o aggregazioni di strati costituenti dei domini di stima; ovviamente affinché queste sub-

popolazioni possano essere qualificate come gruppi è necessario che siano noti i totali di riferimento a livello di ciascuna sottopopolazione.

Le ragioni per le quali si può costruire lo stimatore sotto l'ipotesi che la popolazione sia suddivisa in più gruppi possono essere essenzialmente due:

- i gruppi costituiscono domini d'interesse, per cui si desidera che le stime a livello di gruppo dei totali di alcune variabili ausiliarie (che costituiscono le variabili strutturali della popolazione) coincidano con i totali noti;
- se si suppone che le unità sono relativamente omogenee all'interno dei gruppi, e se esiste una considerevole differenza tra le unità appartenenti a differenti gruppi, allora ha senso introdurre un modello separato per ciascun gruppo, in quanto esso può esprimere la maggior parte della variazione della variabile dipendente y. Ad esempio nel caso in cui si disponga di una unica variabile ausiliaria e si supponga che i rapporti $\frac{y_k}{x_k}$ siano approssimativamente costanti a livello di gruppo e variabili tra i gruppi, ha senso introdurre un modello di regressione del tipo

$$y_k = \beta_g x_k + \varepsilon_k ,$$

con

$$E_\xi(\varepsilon_k) = 0 ; Var_\xi(\varepsilon_k) = c_k\sigma^2 ; \ Cov_\xi(\varepsilon_k, \varepsilon_{k'}) = 0 \text{ per } k \neq k' .$$

L'ipotesi alla base del precedente modello può essere verificata mediante le tecniche usuali di analisi della varianza.

Esempio

Si consideri una popolazione di individui raggruppata in G gruppi $U_{(1)},...,$ $U_{(g)},...,U_{(G)}$ contenenti rispettivamente $N_{(1)},...,$ $N_{(g)},....N_{(G)}$ individui, dove i gruppi sono definiti in base alle modalità incrociate del sesso e delle classi di età.

Si supponga, inoltre, di disporre, per ciascuna unità k, di una variabile ausiliaria x_k di cui sono noti i valori del totale per ciascun gruppo

$$X_{(g)} = \sum_{k=1}^{N(g)} x_k \ .$$

Sotto il modello

$$y_k = \beta_g x_k + \varepsilon_k \ , \qquad \text{per } k \in U_{(g)}$$

con

$$E_\xi(\varepsilon_k) = 0 \ ; Var_\xi(\varepsilon_k) = x_k \sigma^2 \ ; \ Cov_\xi(\varepsilon_k, \varepsilon_{k'}) = 0 \ \text{ per } k \neq k' \ ,$$

sulla base delle espressioni (22) e (23), lo stimatore di regressione generalizzato del totale

$$Y_{(g)} = \sum_{k=1}^{N(g)} y_k$$

a livello di gruppo è uguale a

$$\tilde{Y}_{(g)REG} = \sum_{k \in s} y_k d_k \gamma_k$$

$$= \sum_{k=1}^{n(g)} y_k d_k \left(1 + (X_{(g)} - \sum_{k=1}^{n(g)} d_k x_k) \left(\sum_{k=1}^{n(g)} \frac{d_k x_k^2}{x_k} \right)^{-1} \frac{x_k}{x_k} \right)$$

$$= \sum_{k=1}^{n(g)} y_k d_k \left(1 + (X_{(g)} - \tilde{X}_{(g)})\tilde{X}_{(g)}^{-1}\right)$$

$$= \sum_{k=1}^{n(g)} y_k d_k \frac{X_{(g)}}{\tilde{X}_{(g)}} .$$

Lo stimatore del totale Y è, pertanto, dato da

$$\tilde{Y}_{REG} = \sum_{g=1}^{G} \sum_{k=1}^{n(g)} y_k d_k \frac{X_{(g)}}{\tilde{X}_{(g)}} .$$

che costituisce lo *stimatore del rapporto post-stratificato.* Nel caso in cui, per le unità appartenenti al generico gruppo g (g=1,…,G), x_k è uguale a $\delta_{(g)k}$ si ottiene l'espressione classica *dello stimatore del rapporto post-stratificato* definita da:

$$\tilde{Y}_{REG} = \sum_{g=1}^{G} \sum_{k=1}^{n(g)} y_k d_k \frac{N_{(g)}}{\tilde{N}_{(g)}} ,$$

dove

$$\tilde{N}_{(g)} = \sum_{k=1}^{n(g)} \delta_{(g)k} \, d_k .$$

2.4.2. Modello a livello di grappolo

Introduciamo ora un tipo di stimatore molto interessante dal punto di vista applicativo in quanto viene correntemente utilizzato nelle indagini ISTAT sulle famiglie. Consideriamo, a tal fine una popolazione U di N elementi ripartita in N_I grappoli e con riferimento al grappolo i-esimo $i = \{1,..., N_I\}$ indichiamo con $\mathbf{x}_{Ii}$ il vettore di J variabili ausiliarie; N_{Ii} il numero di unità elementari; $y_{Ii} = \sum_{k=1}^{N_{Ii}} y_{Iik}$ il valore della variabile d'interesse *y*, essendo y_{Iik} il valore della variabile d'interesse *y* della k-esima unità elementare $(k = 1,..., N_{Ii})$ del grappolo. Supponiamo, inoltre, che la popolazione U_I dei grappoli sia suddivisa in G gruppi distinti che definiscono una partizione completa della popolazione stessa. Con riferimento al gruppo g-esimo (g=1,...,G), denotiamo con $U_{I(g)} = \{1,..., i,..., N_{I(g)}\}$ la popolazione dei grappoli e con

$$\mathbf{X}_{I(g)} = \sum_{i=1}^{N_{I(g)}} \mathbf{x}_{Ii}$$

il vettore (supposto noto) dei totali delle J variabili ausiliarie.

Ipotizziamo, quindi, di avere estratto da U_I un campione casuale mediante il seguente schema:

(i) si seleziona un campione $s_I = \{1,..., i,..., n_I\}$ di n_I grappoli mediante il disegno di campionamento che genera l'universo dei campioni S_I e assegna al generico campione s_I la probabilità $p_I(s_I)$ di essere estratto (dove $\sum_{s_I \in S_I} p(s_I) = 1$); di conseguenza, indicando con $S_I(i)$ il sottoinsieme di S_I formato dai campioni contenenti il grappolo i-esimo la probabilità d'inclusione di tale grappolo è data da $\pi_{Ii} = \sum_{s_I \in S(i)} p_I(s_I)$;

(ii) tutte le unità elementari dei grappoli selezionati vengono incluse nel campione; tale circostanza determina il fatto che la probabilità d'inclusione delle unità elementari coincide con quella dei grappoli di appartenenza.

Supponiamo, infine, che il campione $s_{I(g)}$ del gruppo g -definito come $s_{I(g)} = s_I \cap U_{I(g)}$- sia costituito da un numero $n_{I(g)}$ di unità elementari sempre maggiore del numero J di totali noti. Valendo le precedenti condizioni è possibile stimare il totale

$$Y = \sum_{g=1}^{G} \sum_{i=1}^{N_I} \sum_{k=1}^{N_{Ii}} y_{Iik}$$

definendo un modello separato per i grappoli di ciascun gruppo:

$$y_{Ii} = \mathbf{x}_{Ii}' \boldsymbol{\beta}_{I(g)} + \varepsilon_{Ii} \qquad \text{per} \quad i \in U_{I(g)} \qquad (46)$$

dove $\boldsymbol{\beta}_{I(g)}$ è il vettore dei coefficienti di regressione incogniti ed ε_{Ii} è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello sono definiti rispettivamente da

$$E_\xi(\varepsilon_{Ii}) = 0\,,\ \mathrm{Var}_\xi(\varepsilon_{Ii}) = c_{Ii}\sigma^2\,,\ \mathrm{Cov}_\xi(\varepsilon_{Ii}, \varepsilon_{Ii'}) = 0 \ \text{per} \ \forall\, i \neq i'\,; \qquad (47)$$

essendo c_{Ii} delle costanti note.

In base al modello appena introdotto è possibile derivare lo stimatore di regressione generalizzato

$$\tilde{Y}_{REG} = \sum_{g=1}^{G} \sum_{i=1}^{N_{I(g)}} \gamma_{Ii}\, d_{Ii}\, y_{Ii} = \sum_{g=1}^{G} \sum_{i=1}^{N_{I(g)}} w_{Ii}\, y_{Ii}$$

$$= \sum_{g=1}^{G} \sum_{i=1}^{N_{I(g)}} \sum_{k=1}^{N_{Ii}} \gamma_{Ii}\, d_{Ii}\, y_{Iik} = \sum_{g=1}^{G} \sum_{i=1}^{N_{I(g)}} \sum_{k=1}^{N_{Ii}} w_{Ii}\, y_{Iik} \qquad (48)$$

in cui si è indicato con:

$w_{Ii} = d_{Ii}\, \gamma_{Ii}$, il *peso finale*,

$d_{Ii} = \frac{1}{\pi_{Ii}}$, il *peso diretto*,

$\gamma_{Ii} = 1 + (\mathbf{X}_{I(g)} - \tilde{\mathbf{X}}_{I(g)})' \left(\sum_{i=1}^{n_{I(g)}} \frac{d_{Ii}\mathbf{x}_{Ii}\mathbf{x}_{Ii}'}{c_{Ii}} \right)^{-1} \frac{\mathbf{x}_{Ii}}{c_{Ii}}$, il *fattore correttivo del base*, di un campionaria apparte al gruppo g,

$(i \in s_{I(g)})$

essendo

$$\tilde{\mathbf{X}}_{I(g)} = \sum_{i=1}^{n_{I(g)}} \mathbf{x}_{Ii} d_{Ii} .$$

2.5. Tipo di modello

La definizione del *tipo di modello* consiste nella individuazione del modello di regressione, scegliendo in modo opportuno le variabili ausiliarie e le costanti (c_k per i modelli a livello di elemento o c_{Ii} per i modelli a livello di grappolo) che specificano la variabilità dei residui. Dalla definizione congiunta del *tipo,* del *gruppo* e del *livello del modello* è possibile fare discendere i più importanti

stimatori utilizzati nelle indagini campionarie su larga scala. Per illustrare questo aspetto, nel presente paragrafo prenderemo in esame a scopo didattico gli stimatori: *diretto*, *rapporto semplice*, *rapporto post-stratificato*, *ratio-raking* che adottano un modello a livello di unità elementare e che possono essere ottenuti come caso particolare a partire dall'espressione generale dello stimatore di regressione:

$$\tilde{Y}_{REG} = \sum_{k \in s} y_k d_k \gamma_k \,, \qquad (49)$$

definendo in modo opportuno i valori dei correttori γ_k dei pesi base. Altri stimatori, di tipo più complesso, che adottano un modello a livello di grappolo o di unità primaria, sono descritti nei precedenti paragrafi *2.3 e 2.4*.

Stimatore diretto

Definiamo questo tipo di stimatore prendendo in esame un modello a livello di unità elementare ed esaminiamo la situazione in cui è definito un unico gruppo di riferimento costituito dall'intera popolazione. Consideriamo, adesso, un modello di regressione del tipo (4) e (5) in cui:

1. per la generica unità k-esima il vettore delle variabili ausiliarie contiene un solo elemento che assume valore uguale alla probabilità d'inclusione π_k ; inoltre la costante c_k è uguale a π_k ;
2. il vettore dei totali noti delle variabili ausiliarie è costituito da un solo elemento ed è dato da

$$\mathbf{X} = \sum_{k \in U} \pi_k = n \,.$$

Il *tipo di modello* prescelto è definito nel punto 1. e viene formalizzato attraverso la seguente uguaglianza $\mathbf{x}_k = \pi_k = c_k$.

Introducendo le precedenti condizioni nell'espressione del fattore correttivo del peso base definita dalla (23) si ottiene

$$\gamma_k = 1 + (n - \sum_{k \in s} d_k \pi_k) \left(\sum_{k \in s} d_k \pi_k \right)^{-1} \frac{\pi_k}{\pi_k}$$

$$= 1 + \frac{(n - n)}{n} = 1 \qquad (50)$$

Sostituendo[12] , infine, l'espressione di γ_k , appena ottenuta, nella (49) si ottiene la ben nota espressione dello stimatore diretto

$$\tilde{Y} = \sum_{k \in s} y_k d_k \qquad (51)$$

Stimatore rapporto semplice

Definiamo questo tipo di stimatore prendendo in esame un modello a livello di unità elementare ed esaminiamo la situazione in cui è definito un unico gruppo di riferimento costituito dall'intera popolazione. Consideriamo, adesso, un modello di regressione del tipo (4) e (5) in cui:

1. per la generica unità k-esima il vettore delle variabili ausiliarie contiene una sola variabile x_k che assume sempre valori positivi; inoltre la costante c_k è uguale a x_k ;
2. il vettore dei totali noti delle variabili ausiliarie è costituito da un solo elemento ed è dato da

$$\mathbf{X} = \sum_{k \in U} x_k = X .$$

---

12 E' chiaro che il fattori correttivi sono pari a 1 solo nel caso in cui tutte le n unità del campione sono rispondenti all'indagine. Invece, nel caso in cui sono rispondenti all'indagine solamente $n^e < n$ unità campionarie per ottenere lo stimatore diretto occorre utilizzare il seguente tipo di modello alternativo $\mathbf{x_k} = c_k = \pi_k(n/n^e)$.

Il *tipo di modello* prescelto è definito nel punto 1 e viene formalizzato attraverso la seguente uguaglianza $\mathbf{x}_k = x_k = c_k$.

Introducendo le precedenti condizioni nell'espressione del fattore correttivo del peso base definita dalla (23) si ottiene

$$\gamma_k = 1 + (X - \tilde{X})\left(\sum_{k\in s} \frac{d_k x_k^2}{x_k}\right)^{-1} \frac{x_k}{x_k}$$

$$= 1 + (X - \tilde{X})\left(\sum_{k\in s} d_k x_k\right)^{-1}$$

$$= 1 + \frac{(X - \tilde{X})}{\tilde{X}} = \frac{X}{\tilde{X}}. \qquad (52)$$

Sostituendo, infine, l'espressione di γ_k, appena ottenuta, nella (49) si ottiene la ben nota espressione dello stimatore rapporto

$$\tilde{Y} = \frac{\sum_{k\in s} y_k d_k}{\sum_{k\in s} x_k d_k} X = \frac{\tilde{Y}}{\tilde{X}} X. \qquad (53)$$

Stimatore rapporto post-stratificato

Definiamo questo tipo di stimatore prendendo in esame un modello a livello di unità elementare e supponiamo che la popolazione U di elementi sia suddivisa in G gruppi $U_{(1)},...,U_{(g)},...,U_{(G)}$ che definiscono una partizione completa della stessa. Ipotizziamo, inoltre, che siano noti i totali, $X_{(1)},..., X_{(g)},...,X_{(G)}$, di una variabile ausiliaria *x* per tutti i gruppi della partizione. Consideriamo, adesso, un modello di regressione in cui i gruppi di riferimento sono costituiti dalle G sottopopolazioni ed in cui valgono le seguenti condizioni:

1. per la generica unità k-esima il vettore delle variabili ausiliarie contiene una sola variabile x_k che assume sempre valori positivi; inoltre la costante c_k è uguale a x_k ;

2. per ciascuno gruppo g il vettore dei totali noti delle variabili ausiliarie è costituito da un solo elemento, dato da

$$\mathbf{X}_{(g)} = \sum_{k \in U_{(g)}} x_k = X_{(g)} .$$

Introducendo le precedenti condizioni nell'espressione del fattore correttivo del peso base della generica unità $k \in s_{(g)}$ del gruppo g, sotto l'ipotesi che la dimensione $n_{(g)}$ del campione del gruppo g sia maggiore di zero, si ottiene:

$$\gamma_k = 1 + (X_{(g)} - \sum_{k=1}^{n(g)} d_k x_k) \left(\sum_{k=1}^{n(g)} \frac{d_k x_k^2}{x_k} \right)^{-1} \frac{x_k}{x_k}$$

$$= \left(1 + (X_{(g)} - \tilde{X}_{(g)}) \tilde{X}_{(g)}^{-1}\right)$$

$$= \frac{X_{(g)}}{\tilde{X}_{(g)}} \qquad \text{per } (k \in s_{(g)}). \qquad (54)$$

Sostituendo, infine, l'espressione di γ_k , appena ottenuta, nella (49) si ottiene la ben nota espressione dello stimatore rapporto post-stratificato

$$\tilde{Y} = \sum_{g=1}^{G} \frac{\tilde{Y}_{(g)}}{\tilde{X}_{(g)}} X_{(g)} , \qquad (55)$$

in cui

$$\tilde{Y}_{(g)} = \sum_{k=1}^{n(g)} y_k d_k \, .$$

Facciamo notare che lo stimatore[13] (51) definisce come caso particolare tutta una serie di stimatori ben noti nella letteratura sul campionamento, infatti a seconda di come vengono formati i gruppi si hanno:

- *lo stimatore rapporto semplice*, nel caso in cui tutta la popolazione definisca un unico gruppo;
- *lo stimatore del rapporto separato*, nel caso in cui ciascun gruppo sia costituito da un unico strato;
- *lo stimatore del rapporto combinato* nel caso in cui i gruppi siano costruiti come aggregazione di strati;
- *lo stimatore del rapporto post-stratificato,* nel caso in cui le G sottopopolazioni che costituiscono i gruppi siano *post-strati*, in quanto si assume che la variabile utilizzata per definire la partizione in gruppi non sia stata usata per la stratificazione delle unità, ma venga rilevata per ciascuna unità elementare inclusa nel campione; ciò implica, in particolare, che il numero di unità campionarie ricadenti in ciascun *post-strato* è una variabile casuale e ciascun post-strato è costituito dall'unione di parti di strati del disegno di campionamento.

Stimatore ratio-raking

Definiamo questo tipo di stimatore prendendo in esame un modello a livello di unità elementari ed esaminiamo la situazione di una popolazione suddivisa in G gruppi in cui per il generico gruppo g (g=1,...,G) si possano individuare due partizioni distinte. La prima partizione composta di R sottopopolazioni definite sulla base delle modalità assunte dalla variabile x_1, mentre la seconda partizione

---

13 E' utile osservare che lo stimatore espresso dalla (41) può essere ottenuto in modo alternativo a quanto appena fatto utilizzando il modello definito dalle espressioni (41)-(43) in cui tutta la popolazione costituisce un unico gruppo.

è composta di C sottopopolazioni definita sulla base delle modalità assunte dalla variabile x_2. Per ciascun gruppo g (g=1,...,G), il numero di elementi della popolazione appartenenti alla sottopopolazione r (r=1,...,R) della prima partizione è indicato con $N_{(g),1r}$; mentre, si denota con $N_{(g),2c}$ il numero di elementi della popolazione appartenenti alla sottopopolazione c (c=1,...,C) della seconda partizione; supponiamo inoltre che le quantità $N_{(g),1r}$ e $N_{(g),2c}$ siano note.

I dati del problema possono essere riassunti nel modo seguente:

- per ciascuno gruppo g (g=1,...,G), si definisce un vettore di totali noti contenente R+C frequenze assolute:

$$\underline{X}_d^{'} = \left(N_{d,(1,1)},...,N_{d,(1,r)},...,N_{d,(1,R)},N_{d,(2,1)},...,N_{d,(2,c)},N_{d,(2,C)}\right)'$$

- per la generica unità k-esima la costante c_k viene posta uguale ad 1 e si definisce il vettore di variabili ausiliarie, composto di R+C variabili indicatrici:

$$\mathbf{x}_k^{'} = \left(\delta_{k,11},...,\delta_{k,1r},...,\delta_{k,1R},\delta_{k,21},...,\delta_{k,2c},...,\delta_{k,2C}\right)$$

dove $\delta_{k,1r}$ è una variabile indicatrice che assume valore 1 se l'unità k-esima appartiene alla r-esima sottopopolazione della prima partizione e valore 0 altrimenti (r=1,...,R); $\delta_{k,2c}$ è una variabile indicatrice che assume valore 1 se l'unità k-esima appartiene alla c-esima sottopopolazione della seconda partizione e valore 0 altrimenti (c=1,...,C).

Introducendo le precedenti condizioni nell'espressione del fattore correttivo del peso base della generica unità $k \in s_{(g)}$ del gruppo g appartenente alla r-esima sottopopolazione della prima partizione ed alla c-esima sottopopolazione della seconda partizione si ottiene:

$$\gamma_k = 1 + \left[N_{(g),11} - \tilde{N}_{(g),11}, ..., N_{(g),1r}, -\tilde{N}_{(g),1r}, ..., N_{(g),2c}, -\tilde{N}_{(g),2c}, ..., N_{(g),2C}, -\tilde{N}_{(g),2C} \right]$$

$$\times \begin{bmatrix} \mathbf{A}_{RR} & \mathbf{A}_{RC} \\ \mathbf{A}'_{RC} & \mathbf{A}_{CC} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_{r-1} \\ 1 \\ \mathbf{0}_{R+c-r-1} \\ 1 \\ \mathbf{0}_{C-c} \end{bmatrix} \qquad \text{(per } (k \in s_{(g)}) \cap (\delta_{k,1r}\delta_{k,2c} = 1)\text{)},$$

(56)

dove abbiamo indicato con: $\times$ l'operatore di prodotto matriciale; $\mathbf{0}_\nu$ un vettore costituito da ν valori identicamente pari a zero; $\mathbf{A}_{RR}$ una matrice diagonale di dimensione (R x R) il cui i-esimo (i=1,...,R) elemento sulla diagonale principale è dato da $\tilde{N}_{(g),1i} = \sum_{k \in s_{(g)}} d_k \delta_{k,1i}$; $\mathbf{A}_{RC}$ una matrice di dimensione (R x C) il cui elemento che occupa la riga i-esima (i=1,...,R) e la colonna j-esima (j=1,...,C) è espresso da $\tilde{N}_{(g),1i,2j} = \sum_{k \in s_{(g)}} d_k \delta_{k,1i} \delta_{k,2j}$; $\mathbf{A}_{CC}$ una matrice diagonale di dimensione (C x C) il cui j-esimo (j=1,...,C) elemento sulla diagonale principale è calcolato come $\tilde{N}_{(g),2c} = \sum_{k \in s_{(g)}} d_k \delta_{k,2j}$.

Dopo alcuni passaggi, indicando con

$$\begin{bmatrix} \mathbf{A}_{RR} & \mathbf{A}_{RC} \\ \mathbf{A}'_{RC} & \mathbf{A}_{CC} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}_{RR} & \mathbf{B}_{RC} \\ \mathbf{B}'_{RC} & \mathbf{B}_{CC} \end{bmatrix},$$

ed utilizzando i risultati standard sulle inverse delle matrici a blocchi (Searle, 1971, pag. 27) si ottiene che il fattore correttivo del peso base della generica unità $k \in s_{(g)}$ del gruppo g appartenente alla r-esima sottopopolazione

della prima partizione ed alla c-esima sottopopolazione della seconda partizione è espresso da

$$\gamma_k = 1 + \sum_{i=1}^{R} \frac{N_{(g),1i} - \tilde{N}_{(g),1i}}{b_{(RR),ir} + b_{(RC),ic}} + \sum_{j=1}^{C} \frac{N_{(g),2j} - \tilde{N}_{(g),2j}}{b_{(CC),jc} + b_{(RC),rj}},$$

dove abbiamo indicato con: $b_{(RR),ir}$ l'elemento nella riga i-esima e nella colonna r-esima

della matrice $\mathbf{B}_{RR} = \mathbf{A}_{RR}^{-1} + \mathbf{A}_{RR}^{-1}\mathbf{A}_{RC}\left(\mathbf{A}_{CC} - \mathbf{A}_{RC}^{'}\mathbf{A}_{RR}^{-1}\mathbf{A}_{RC}\right)^{-1}\mathbf{A}_{RC}^{'}\mathbf{A}_{RR}^{-1}$; $b_{(RC),ic}$ l'elemento nella riga i-esima e nella colonna c-esima della matrice $\mathbf{B}_{RC} = -\mathbf{A}_{RR}^{-1}\mathbf{A}_{RC}\left(\mathbf{A}_{CC} - \mathbf{A}_{RC}^{'}\mathbf{A}_{RR}^{-1}\mathbf{A}_{RC}\right)^{-1}$; $b_{(CC),jc}$ l'elemento nella riga j-esima e nella colonna r-esima della matrice $\mathbf{B}_{CC} = -\left(\mathbf{A}_{CC} - \mathbf{A}_{RC}^{'}\mathbf{A}_{RR}^{-1}\mathbf{A}_{RC}\right)^{-1}\mathbf{A}_{RC}^{'}\mathbf{A}_{RR}^{-1}$; $b_{(RC),rj}$ l'elemento nella riga r-esima e nella colonna j-esima della matrice $\mathbf{B}_{RC}$.

BIBLIOGRAFIA

Brewer, K.R.V., Hanif, M., 1983, Sampling with Unequal Probabilities, Springer-Verlag. New-York.

Chen, P. P. S., 1976, *The Entity-Relationship Model. Towards a Unified View of Data*, ACM Trans. Database System 1, n. 1.

Cochran, W. G., 1977, Sampling Techniques, Wiley, New York.

Deville, J. C., Särndal, C. E., 1992, *Calibration Estimators in Survey Sampling,* Journal of the American Statistical Association, vol. 87, pp. 367-382.

De Vitiis, C., Pagliuca, D., 2003, *La presentazione sintetica degli errori campionari e l'analisi grafica degli outlier nel software Genesees*, Atti del Convegno Intermedio "Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia" della Società Italiana di Statistica (su CD-ROM).

Falorsi, P.D., Falorsi, S., 1995, *Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese*, Rapporto di ricerca CON.PRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, n. 13.

Falorsi, P.D., Falorsi, S., 1997, *The Italian Generalized Package for Weighting Persons and Families: Some Experimental Results with Different Non-Response Models*, Statistics in Transitions Journal of the Polish Statistical Association, vol. 3, n. 2.

Falorsi, P. D., Falorsi S., 1998, *The Italian generalized estimation package: some experimental results for estimation on households suveys with different non response mechanism*, Quaderni di Ricerca, ISTAT, n.4, pp.63-94.

Falorsi, S., Rinaldelli, C., 1998, *Un Software generalizzato per il calcolo delle stime e degli errori di campionamento*, Statistica Applicata, vol**.** 10, n. 2 , pp. 217-234.

Falorsi, S., Pagliuca, D., Scepi, G., 1999, *Generalised Software for Sampling Errors – GSSE"*, Proceedings of the Seminar on Exchange of Technology and Know-How (ETK 99), held in Prague, Czech Republic on the 13-15 October 1999, pp. 169-175.

Falorsi, S., Pagliuca, D., Scepi, G., 2000, *Generalised Software for Sampling Errors – GSSE"*, Research in Official Statistics - ROS, vol. 3, n. 2, pp. 89-108.

Horvitz, D.G., Thompson, D. J, 1952, *A Generalization of Sampling without Replacement from Finite Universe*, Journal of the American Statistical Association, vol. 47, pp. 663-685.

Kish, L., 1965, Survey Sampling, Wiley, New York.

Pagliuca, D. (a cura di), 2004a, *Genesees v.3.0., Funzione Stime ed Errori* Manuale utente ed aspetti metodologici, Tecniche e Strumenti, ISTAT, n. 3.

Pagliuca, D. (a cura di)*,* 2004b, *Genesees v.3.0.*, *Funzione Analisi dei Modelli* Manuale utente ed aspetti metodologici, Tecniche e Strumenti, ISTAT, n. 4.

Russo A., 1987, *Sulla Presentazione degli Errori di Campionamento mediante Modelli. Il Metodo dei Modelli Regressivi*, Quaderni di Discussione, ISTAT, n. 87, 04.

Särndal, C.E., Swensson , B. and Wretman, J., 1989, *The weighted residual technique for estimating the variance of the general regression estimator of the finite population total*, Biometrika, vol. 76, n. 3, pp. 527-537

Särndal, C.E., Swensson, B. and Wretman, J., 1992, Model Assisted Survey Sampling, Springer-Verlag. New-York.

Singh, A. C., Mohl, C. A., 1996, *Understanding CalibrationEstimators in Survey Sampling*, Survey Methodology, vol. 22, n. 2, pp. 107-115.

Verma, V., Scott, C., O'Muircheartaigh, C., 1980, *Sample Designs and Sampling Errors fo the Word Fertility Survey*, Journal of the Royal Statistical Society A, vol. 143,Part. 4, pp. 431-473.

Verma, V., 1982, *The Estimation and Presentation of Sampling Errors*, Technical Bulletins, World Fertility Survey, New York.

Wolter, K. M., 1985 Introduction to variance estimation. Springer-Verlag. New York.

Woodruff, R.S., 1971, *A Simple Method for Approximating the Variance of a Complicated Estimate*, Journal of the American Statistical Association, vol.66, n. 334, pp. 411-414.