Sistema statistico nazionale

Istituto nazionale di statistica

Stime ed Errori

Note metodologiche

2005

A cura di:

Paragrafi: 1 Paolo Righi; **2** Fabrizio Solari; **3** e **4** Stefano Falorsi.

Indice

1 CENNI SULLA DEFINIZIONE DELLO STIMATORE DI REGRESSIONE GENERALIZZATA

Per descrivere la metodologia adottata per il calcolo degli errori di campionamento per la stima di un totale, si prenda in considerazione una popolazione $U = \{1, \ldots, k, \ldots, N\}$, di N elementi, e si denoti con Y la variabile oggetto d'indagine. Sia quindi

$$Y = \sum_{k \in U} y_k$$

il parametro da stimare, essendo y_k il valore della variabile d'interesse Y assunto dalla generica unità k.

Descriviamo come calcolare gli errori campionari di un ampia classe di stimatori diretti di Y, i quali possono essere derivati dalla teoria degli stimatori di regressione generalizzata. Tali stimatori appartengono, a loro volta, alla classe degli stimatori di calibrazione che, in estrema sintesi, definiscono i coefficienti finali delle unità attraverso la risoluzione di un problema di minimo vincolato. In particolare, dati dei totali noti a livello di popolazione (o sottopopolazione), per alcune variabili ausiliarie il processo di ottimizzazione avviene minimizzando la distanza tra i coefficienti diretti (pari all'inverso della probabilità di inclusione nel campione), eventualmente corretti in presenza di mancate risposte totali, e i coefficienti finali (incogniti) assegnati alle unità campionarie, con il vincolo che le stime ottenute con i coefficienti finali riproducano i totali noti sopra definiti.

Ciascuno stimatore di calibrazione si distingue sia per il tipo di totali noti utilizzati che per altri due elementi riguardanti: la funzione di distanza impiegata, per valutare lo scostamento tra i coefficienti diretti e quelli finali; il peso, c_k, attribuito a ciascuna unità del campione, che interviene come fattore moltiplicativo della distanza calcolata tra coefficiente diretto e finale per l'unità k-esima.

Si dimostra che gli stimatori di regressione generalizzata sono un caso particolare degli stimatori di calibrazione, quando la distanza scelta per l'ottenimento dei pesi finali è quella euclidea (Deville e Särndal 1992).

In tal caso,con riferimento ad un campione casuale s = {1, …, k, …, n} di n unità, il problema di minimo vincolato è rappresentato dal seguente sistema

$$\begin{cases} min\left[\sum_{k\in s}\dfrac{(1/\pi_k - w_k)^2}{1/\pi_k}\cdot c_k\right], \\ \sum_{k\in s} w_k \boldsymbol{x}_k = \boldsymbol{X} \end{cases}$$

in cui, relativamente alla k-esima unità appartenente al campione, si ha che nella prima espressione (funzione obiettivo) π_k è la probabilità di inclusione, w_k è il peso finale calibrato incognito e c_k è un peso indipendente da π_k attribuito a ciascuna unità del campione.

Nella seconda espressione, detta *equazione di calibrazione*, sono contenuti i vincoli e $\boldsymbol{x}_k = (x_{1k},\ldots,x_{jk},\ldots,x_{Jk})'$ rappresenta il vettore dei valori assunti dalle J variabili ausiliarie **X**= (X_1, …, X_j, …X_J) per le quali sono noti i totali $\boldsymbol{X} = (X_1,\ldots,X_j,\ldots X_J)'$ riferiti all'intera popolazione (o eventualmente a particolari sottopopolazioni).

Un importante risultato ottenuto in Deville e Särndal (1992) indica che, nelle indagini su larga scala, gli stimatori di calibrazione che utilizzano una generica funzione di distanza sono asintoticamente equivalenti ai corrispondenti stimatori di regressione generalizzata che usano la distanza euclidea[1]. Alla luce di questo risultato la stima della varianza di tutti gli stimatori di calibrazione può essere approssimata dalla stima della varianza calcolata sui corrispondenti stimatori di regressione per i quali è possibile derivare l'espressione esplicita della stima della varianza.

Restringendo pertanto l'attenzione alla classe degli stimatori di regressione generalizzata, secondo una trattazione generale questi si fondano sulle seguenti informazioni:

---

1. Più precisamente per assicurare l'equivalenza asintotica fra le stime prodotte con uno stimatore di calibrazione e quelle prodotte con uno stimatore di regressione generalizzata, la funzione di distanza del primo stimatore deve rispettare alcune deboli condizioni (Deville e Särndal 1992).

- per ciascun elemento del campione *k* si conosce il vettore delle *J*+1 osservazioni $(y_k, \boldsymbol{x}_k)$, in cui $\boldsymbol{x}_k = (x_{1k},\ldots,x_{jk},\ldots,x_{Jk})'$ è il vettore dei valori assunti dalle *J* variabili ausiliarie **X**= (X_1, …, X_j, …X_J) ;
- risulta noto il vettore dei totali $\boldsymbol{X} = (X_1,\ldots,X_j,\ldots X_J)'$ corrispondenti alle *J* variabili ausiliarie.

Lo stimatore in questione sfrutta le suddette informazioni ausiliarie attraverso la definizione di un modello di regressione lineare ξ che spiega la nuvola dei punti individuata dall'insieme $\{(y_k, \boldsymbol{x}_k): k = 1,...,N\}$. Il modello si basa sulle seguenti ipotesi:

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 al vettore $\boldsymbol{x}_k$ (*k*=1,.....,*N*) è la seguente:

$$y_k = \boldsymbol{x}_k'\boldsymbol{\beta} + \varepsilon_k \quad (k=1...,N) \qquad (1.1)$$

in cui $\boldsymbol{\beta} = (\beta_1,...,\beta_j,...,\beta_J)'$ 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:

$$E_\xi(\varepsilon_k) = 0,\ Var_\xi(\varepsilon_k) = c_k\sigma^2,\ Cov_\xi(\varepsilon_k,\varepsilon_l) = 0 \quad \text{per} \forall\, k \neq l; \qquad (1.2)$$

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

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.

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 $\boldsymbol{\beta}$ del modello ξ. Utilizzando la teoria standard della regressione generalizzata, si ha che il miglior stimatore lineare non distorto dei coefficienti $\boldsymbol{\beta}$, sotto il modello ξ, è dato da:

$$\boldsymbol{B} = \left(B_1,...,B_j,...,B_J\right)' = \left(\sum_{k\in U}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{c_k}\right)^{-1}\sum_{k\in U}\frac{\boldsymbol{x}_k y_k}{c_k}. \qquad (1.3)$$

Il vettore dei coefficienti $\boldsymbol{B}$ è, tuttavia, una caratteristica incognita della popolazione in quanto le variabili **X** e Y non sono note per l'intero universo. Si può, pertanto, procedere ad una stima di $\boldsymbol{B}$ mediante i dati rilevati sul campione *s*. Poiché la relazione (1.3) si presenta come il prodotto di una funzione dei totali della popolazione;

$$\boldsymbol{T}_1 = \sum_{k\in U}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{c_k} \qquad \text{e} \qquad \boldsymbol{T}_2 = \sum_{k\in U}\frac{\boldsymbol{x}_k y_k}{c_k},$$

una stima asintoticamente corretta di $\boldsymbol{B}$ può essere ottenuta stimando ciascun totale mediante lo stimatore di Horvitz-Thompson. I due stimatori sono espressi attraverso le seguenti formule

$$\hat{\boldsymbol{T}}_1 = \sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\pi_k c_k} \qquad \text{e} \qquad \hat{\boldsymbol{T}}_2 = \sum_{k\in s}\frac{\boldsymbol{x}_k y_k}{\pi_k c_k}.$$

La stima di **B** assume, pertanto, la seguente forma:

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

Sulla base di $\hat{\boldsymbol{B}}$ è possibile, quindi, calcolare con riferimento alle *N* unità della popolazione, i valori interpolati $\hat{y}_1,...,\hat{y}_k,...,\hat{y}_N$, relativi ai corrispondenti valori $y_1,...,y_k,...,y_N$, mediante la relazione

$$\hat{y}_k = \boldsymbol{x}_k'\hat{\boldsymbol{B}}, \ (k = 1,.....,N) \ . \qquad (1.4)$$

In letteratura sono proposti due approcci per la stima di $\boldsymbol{B}$ che usano alternativamente o i coefficienti diretti o quelli finali di riporto.

Con riferimento alle *n* unità del campione e in base alla (1.4) i residui sono dati da

$$e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}\,,\ (k = 1,.....,n)\,. \qquad (1.5)$$

Per la (1.5), il totale *Y* può, quindi, essere riscritto mediante la seguente espressione

$$Y = \sum_{k \in U} y_k = \sum_{k \in U} \hat{y}_k + \sum_{k \in U} e_k\ . \qquad (1.6)$$

Dalla (1.6) si osserva 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 $\hat{y}_k$ può essere definito per tutte le unità della popolazione; il secondo, invece, rappresenta una quantità incognita, poiché è possibile calcolare i residui e_k solo per le unità appartenenti al campione osservato. Sostituendo quindi nella (1.6) lo stimatore di Horvitz-Thompson di tale totale incognito, si ottiene lo stimatore di regressione generalizzata del totale *Y*

$$\hat{Y}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} \frac{e_k}{\pi_k}\,. \qquad (1.7)$$

Considerando che il termine $\sum_{k \in U} \hat{y}_k$ si può riformulare come

$$\sum_{k \in U} \hat{y}_k = \sum_{k \in U} \boldsymbol{x}_k' \hat{\boldsymbol{B}} = \left(\sum_{k \in U} \boldsymbol{x}_k \right)' \hat{\boldsymbol{B}} = \boldsymbol{X}' \hat{\boldsymbol{B}} \qquad (1.8)$$

e che il secondo totale delle (1.7) può essere riscritto mediante il seguente passaggio

$$\sum_{k \in s} \frac{e_k}{\pi_k} = \sum_{k \in s} \frac{\left(y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}} \right)}{\pi_k} = \sum_{k \in s} \left(\frac{y_k}{\pi_k} \right) - \sum_{k \in s} \left(\frac{\boldsymbol{x}_k}{\pi_k} \right)' \hat{\boldsymbol{B}} = \hat{Y} - \hat{\boldsymbol{X}}' \hat{\boldsymbol{B}}\,, \qquad (1.9)$$

in cui $\hat{Y}$ e $\hat{\boldsymbol{X}}$ indicano le stime di Horvitz-Thompson dei corrispondenti totali *Y* e ***X***, è possibile riformulare la (1.7) secondo l'espressione

$$\hat{Y}_{GREG} = \hat{Y} + (\boldsymbol{X} - \hat{\boldsymbol{X}})' \hat{\boldsymbol{B}} \qquad (1.10)$$

dalla quale risulta che lo stimatore di regressione generalizzata è ottenuto come somma dello stimatore di Horvitz-Thompson del totale *Y* più un termine di aggiustamento regressivo che dipende dalle differenze tra i totali noti ***X*** e le

corrispondenti stime campionarie di Horvitz-Thompson $\hat{\boldsymbol{X}}$ ponderate con i rispettivi coefficienti di regressione stimati $\hat{\boldsymbol{B}}$.

Dalla (1.10), attraverso alcuni semplici passaggi lo stimatore si può riscrivere come

$$\hat{Y}_{GREG} = \sum_{k \in s} \frac{g_{ks} y_k}{\pi_k} \tag{1.11}$$

dove compare il fattore correttivo del peso diretto $1/\pi_k$:

$$g_{ks} = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})' \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{\pi_k c_k} \right)^{-1} \frac{\boldsymbol{x}_k}{c_k} . \tag{1.12}$$

Una importante proprietà dello stimatore di regressione generalizzata è che la stima dei totali di popolazione delle variabili ausiliarie è uguale ai corrispondenti totali noti. Sostituendo nella (1.11) y_k con $\boldsymbol{x}_k$ si ha, infatti, $\sum_{k \in s} \frac{g_{ks} \boldsymbol{x}_k}{\pi_k} = \boldsymbol{X}$.

Una definizione più precisa dello stimatore di regressione generalizzata passa attraverso l'introduzione di tre concetti che specificano ulteriormente la relazione della variabile d'interesse con il relativo modello di regressione. Questi sono: il *gruppo di riferimento del modello* (*model group*), il *livello del modello* (*model level*) ed il *tipo di modello* (*model type*).

1.1 GRUPPO DI RIFERIMENTO DEL MODELLO

Data una partizione completa della popolazione *U*, $\{U_1, \ldots, U_d, \ldots, U_D\}$, si definisce il generico *gruppo di riferimento del modello* U_d un sottoinsieme (o sottopopolazione) in cui,

- sono noti i totali di una o più variabili ausiliarie. Occorre notare che non è necessario che l'insieme delle variabili ausiliarie sia lo stesso per ciascuna sottopopolazione.
- il campione s_d appartenente al gruppo di riferimento *d,* definito come $s_d = s \cap U_d$, deve essere sempre costituito da un numero di unità maggiore del numero di totali noti.

Valendo le precedenti condizioni è possibile definire un modello separato per le unità di ciascun gruppo. Rispetto alla (1.1), in cui il gruppo di riferimento è l'intero universo *U*, si costruisce quindi un modello di regressione per ciascun U_d, espresso da

$$y_k = \boldsymbol{x}'_{dk}\boldsymbol{\beta}_d + \varepsilon_k \quad \forall k \in U_d\,, \tag{1.13}$$

in cui valgono le ipotesi (1.2) ed in cui $\boldsymbol{x}_{dk}$ è il vettore dei valori assunti, dall'unità *k,* sulle variabili ausiliarie utilizzate per la costruzione del modello, nella sottopopolazione U_d.

Analogamente alla (1.3) la stima del vettore $\boldsymbol{\beta}_d$ si ottiene come:

$$\hat{\mathbf{B}}_d = \left(\sum_{k \in s_d} \frac{\boldsymbol{x}_{dk}\boldsymbol{x}'_{dk}}{\pi_k c_k} \right)^{-1} \sum_{k \in s_d} \frac{\boldsymbol{x}_{dk} y_k}{\pi_k c_k}\,.$$

Lo stimatore di regressione generalizzata basato su una suddivisione dell'universo in gruppi di riferimento è dato da:

$$\hat{Y}_{GREG} = \sum_{d=1}^{D} \sum_{k \in s_d} \frac{g_{ks_d} y_k}{\pi_k}\,,$$

nella quale

$$g_{ks_d} = 1 + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_d)' \left(\sum_{k \in s_d} \frac{\boldsymbol{x}_{dk} \boldsymbol{x}'_{dk}}{\pi_k c_k} \right)^{-1} \frac{\boldsymbol{x}_{dk}}{c_k} \qquad (1.14)$$

con $\boldsymbol{X}_d = \sum_{U_d} \boldsymbol{x}_{dk}$ e $\hat{\boldsymbol{X}}_d = \sum_{s_d} \boldsymbol{x}_{dk} \big/ \pi_k$.

1.2 LIVELLO DEL MODELLO

Il concetto di livello del modello è relativo al tipo di unità utilizzata nella formulazione del modello. Ad esempio il modello può essere formulato a livello di:

a) *unità elementare*, se nella sua definizione le variabili d'interesse e quelle ausiliarie si riferiscono a ciascuna unità elementare della popolazione;

b) *cluster* (o gruppi) di elementi, se nella sua definizione le variabili d'interesse e quelle ausiliarie si riferiscono a grappoli di unità elementari della popolazione.

In assenza di gruppi di riferimento del modello il caso a) prevede che nella relazione (1.1) e sotto le ipotesi (1.2), *k* indichi la generica unità elementare.
Per il caso b), definito con $U_I = \{1,\ldots,i,\ldots,N_I\}$, l'universo dei *cluster*, si può costruire il seguente modello di regressione ξ_I

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta}_I + \varepsilon_i \ , \qquad (i=1\ldots,N_I) \tag{1.15}$$

in cui

$Y_i = \sum_{k\in i} y_k$ e $\boldsymbol{X}_i = \sum_{k\in i} \boldsymbol{x}_k$ sono i totali di Y e **X** per il generico cluster *i*;

$\boldsymbol{\beta}_I = (\beta_{I1},\ldots,\beta_{Ij},\ldots,\beta_{IJ})'$ è il vettore dei *J* coefficienti di regressione incogniti;

ε_i è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello ξ_I sono definiti rispettivamente da:

$$E_{\xi_I}(\varepsilon_i)=0\,,\ Var_{\xi_I}(\varepsilon_i)=c_i\sigma_I^2\,,\ Cov_{\xi_I}(\varepsilon_i,\varepsilon_{i'})=0 \quad \text{per}\ \forall\, i\neq i'\,; \tag{1.16}$$

essendo le c_i $(\text{per}\ i\in U_I)$ delle costanti note.

Lo stimatore di regressione definito a livello di *cluster* dato dalla (1.15) e dalla (1.16) assume, dunque, la seguente espressione:

$$\hat{Y}_{GREG} = \sum_{i \in s_I} \frac{g_{ks_I} Y_i}{\pi_i}$$

in cui s_I è il campione dei cluster e

$$g_{ks_I} = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})' \left(\sum_{i \in s_I} \frac{\boldsymbol{X}_i \boldsymbol{X}_i'}{\pi_i c_i} \right)^{-1} \frac{\boldsymbol{X}_i}{c_i} \qquad (1.17)$$

è il fattore correttivo del peso diretto e π_i è la probabilità di inclusione del cluster *i* nel campione s_I.

Espressioni analoghe alla (1.13) e alla (1.14) si ottengono quando la popolazione U_I è partizionata in $U_{I1},\ldots,U_{Id},\ldots,U_{ID}$ gruppi di riferimento. La relazione che lega la variabile oggetto d'indagine e le variabili ausiliarie è data da

$$Y_i = \boldsymbol{X}'_{di} \boldsymbol{\beta}_{Id} + \varepsilon_i \quad \forall i \in U_{Id}$$

in cui $\boldsymbol{X}_{di}$ è il vettore dei totali calcolati sul cluster *i* delle variabili ausiliarie utilizzate per la costruzione del modello nella sottopopolazione U_{Id}.

Lo stimatore di regressione si può, pertanto, formulare attraverso la relazione

$$\hat{Y}_{GREG} = \sum_{d=1}^{D} \sum_{k \in s_{Id}} \frac{g_{ks_{Id}} y_k}{\pi_k},$$

in cui $s_{Id} = s_I \cap U_{Id}$;

$$g_{ks_{Id}} = 1 + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_d)' \left(\sum_{i \in s_{Id}} \frac{\boldsymbol{X}_{di} \boldsymbol{X}'_{di}}{\pi_i c_i} \right)^{-1} \frac{\boldsymbol{X}_{di}}{c_i} \qquad (1.18)$$

è il fattore correttivo calcolato a livello di *cluster*.

Si ricorda che un modello a livello di unità elementare corrisponde ad uno stimatore che attribuisce un peso finale diverso per tutte le unità elementari appartenenti ad una medesima unità finale di campionamento; viceversa, un modello a livello di *cluster* di unità elementari corrisponde ad uno stimatore che attribuisce un peso finale uguale per tutte le unità elementari appartenenti ad una medesima unità finale di campionamento.

Infine si ricorda che, mentre per impostare un modello a livello di unità elementare non vi sono vincoli sul tipo di disegno campionario adottato, per definire un modello di regressione a livello di cluster è necessario aver utilizzato un disegno in cui le unità finali di campionamento sono dei grappoli.

1.3 TIPO DI MODELLO

La scelta delle variabili ausiliarie **X** = (X_1, …, X_j, …X_J) e del parametro c_k determina il *tipo di modello* sottostante allo stimatore di regressione generalizzata. In particolare, la specificazione di **X** e c_k, associata alla definizione del livello e del gruppo di riferimento, conducono a noti stimatori che possono essere derivati anche al di fuori della teoria degli stimatori di calibrazione. Nella tabella 1, relativamente a campioni di unità elementari, si descrive il legame esistente tra alcuni degli stimatori più usati in letteratura e la classe degli stimatori di calibrazione.

Tabella 1 – Alcuni casi particolari dello stimatore di calibrazione per campioni di unità elementari

Stimatore	Gruppi di riferimento del modello	Tipo di modello		Fattore correttivo g_{ks}	Forma dello stimatore
		Valori assunti da x_k o x_{dk}	Valori assunti da c_k		
Horvitz-Thompson	No	π_k	π_k	1	$\hat{Y}$
Espansione per disegni semplici	No	n/N	n/N	1	$\hat{Y}_{espansione}$
Hàjek	Totale popolazione	1	1	$N/\hat{N}$	$\frac{\hat{Y}}{\hat{N}}N$
Rapporto semplice	Totale popolazione	x_k	x_k	$X/\hat{X}$	$\frac{\hat{Y}}{\hat{X}}X$
Rapporto separato	Ciascun gruppo coincide con uno strato (*d*≡*h*)	x_{dk}	x_{dk}	$X_h/\hat{X}_h$	$\sum_{h=1}^{H}\frac{\hat{Y}_h}{\hat{X}_h}X_h$
Rapporto combinato	Totale popolazione	x_k	x_k	$X/\sum_h \hat{X}_h$	$\frac{\sum_h \hat{Y}_h}{\sum_h \hat{X}_h}X$
Rapporto combinato per sottopopolazioni	Ciascun gruppo *d* è costruito come aggregazione di strati	x_{dk}	x_{dk}	$X_d/\sum_{h\in d}\hat{X}_h$	$\sum_{d=1}^{D}\frac{\sum_{h\in d}\hat{Y}_h}{\sum_{h\in d}\hat{X}_h}X_d$
Rapporto post-stratificato*	Ciascun gruppo coincide con un post-strato (*d*≡*a*)†	x_{dk}	x_{dk}	${}_aX/{}_a\hat{X}$	$\sum_{a=1}^{A}\frac{{}_a\hat{Y}}{{}_a\hat{X}}{}_aX$
Rapporto post-stratificato separato**	Ciascun gruppo coincide con una combinazione tra post-strato e strato (*d*≡*a*∩*h*)	x_{dk}	x_{dk}	${}_aX_h/{}_a\hat{X}_h$	$\sum_{a=1}^{A}\sum_{h=1}^{H}\frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}{}_aX_h$
Rapporto post-stratificato combinato**	Ciascun gruppo coincide con un post-strato (*d*≡*a*)	x_{dk}	x_{dk}	${}_aX_h/\sum_h {}_a\hat{X}_h$	$\sum_{a=1}^{A}\frac{\sum_h {}_a\hat{Y}_h}{\sum_h {}_a\hat{X}_h}{}_aX$

**Utilizzato con un disegno semplice; ** utilizzato con disegno stratificato; † Il generico post-strato è indicato con a (a=1, ..., A);*

Gli stimatori presentati nella tabella 1 si possono agevolmente estendere ai casi di disegni a grappoli o a due o più stadi di campionamento.

2 LINEARIZZAZIONE DELLO STIMATORE DI REGRESSIONE GENERALIZZATA

Per quanto illustrato nel *paragrafo 1*, una delle possibili espressioni dello stimatore di regressione generalizzata è

$$\hat{Y}_{GREG} = \hat{Y} + (\boldsymbol{X} - \hat{\boldsymbol{X}})'\hat{\boldsymbol{B}},$$

che può essere riscritta nel seguente modo

$$\hat{Y}_{GREG} = \hat{Y} + (\boldsymbol{X} - \hat{\boldsymbol{X}})'\hat{\boldsymbol{T}}_1^{-1}\hat{\boldsymbol{T}}_2 . \qquad (2.1)$$

La (2.1) evidenzia come $\hat{Y}_{GREG}$ sia una funzione non lineare degli stimatori lineari non distorti $\hat{Y}$, $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{T}}_1$ e $\hat{\boldsymbol{T}}_2$ rispettivamente dei totali Y, $\boldsymbol{X}$, $\boldsymbol{T}_1$ e $\boldsymbol{T}_2$.

Sia in generale $\tilde{Y} = f(\hat{\Theta}_1, \ldots, \hat{\Theta}_q)$ uno stimatore del parametro $Y = f(\Theta_1, \ldots, \Theta_q)$, in cui *f* è una funzione non lineare e il generico $\hat{\Theta}_i$ è uno stimatore lineare non distorto del totale Θ_i della variabile ϑ_i, (*i* = 1, …, *q*).

In presenza di funzioni non lineari, si pone il problema della determinazione della stima della media e della varianza di $\tilde{Y}$ .Risolviamo tale problema con il metodo della linearizzazione in serie di Taylor, il quale consiste nell'approssimare lo stimatore $\tilde{Y}$ con una funzione lineare dei $\hat{\Theta}_i$.

Per applicare il metodo è necessario che *f* sia differenziabile almeno fino al secondo ordine in un intorno sufficientemente ampio del punto $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_q)'$. Indicato con $\hat{\boldsymbol{\Theta}}$ il vettore $(\hat{\Theta}_1, \ldots, \hat{\Theta}_q)'$, lo sviluppo in serie di Taylor di $\tilde{Y}$ intorno a $\boldsymbol{\Theta}$ rispetto alle variabili $\hat{\Theta}_i$ porta all'identità

$$\tilde{Y} = f(\boldsymbol{\Theta}) + \sum_{i=1}^{q} g_i(\boldsymbol{\Theta})(\hat{\Theta}_i - \Theta_i) + R_2, \qquad (2.2)$$

dove

$$g_i(\boldsymbol{\Theta}) = \left[\frac{\partial f(\hat{\boldsymbol{\Theta}})}{\partial \hat{\Theta}_i} \right]_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}}$$

è il valore assunto dalla derivata parziale di $\tilde{Y}$ rispetto a $\hat{\Theta}_i$ calcolata nel punto $\boldsymbol{\Theta}$, mentre R_2 è il resto della formula di Taylor, espresso come funzione dei termini di ordine superiore al primo. Se la dimensione campionaria *n* è sufficientemente elevata, R_2 può essere considerato trascurabile rispetto agli altri termini. Quindi, essendo $f(\boldsymbol{\Theta}) = Y$, la (2.2) si può scrivere come

$$\tilde{Y} - Y \doteq \sum_{i=1}^{q} g_i(\boldsymbol{\Theta})(\hat{\Theta}_i - \Theta_i). \qquad (2.3)$$

Calcolando il valore atteso in entrambi i membri, si ottiene

$$E(\tilde{Y}) - Y \doteq \sum_{i=1}^{q} g_i(\boldsymbol{\Theta})\left[E(\hat{\Theta}_i) - \Theta_i\right] = 0,$$

dalla quale si deduce che $\tilde{Y}$ è uno stimatore approssimativamente corretto di *Y*. Di conseguenza, elevando entrambi i membri della (2.3) al quadrato e passando ai valori attesi si ha

$$V(\tilde{Y}) = E(\tilde{Y} - Y)^2 \doteq V\left[\sum_{i=1}^{q} g_i(\boldsymbol{\Theta})\hat{\Theta}_i \right]. \qquad (2.4)$$

La (2.4) richiede il calcolo delle varianze e covarianze degli stimatori $\hat{\Theta}_i$, operazione che dal punto di vista computazionale può risultare piuttosto onerosa. Per ovviare a tale inconveniente, è possibile ricorrere alla trasformata di Woodruff (1971). Infatti, l'approssimazione della varianza di $\tilde{Y}$ data dalla (2.4) si può riformulare mediante la varianza dello stimatore corretto del totale

$$Z = \sum_{k \in U} z_k$$

in cui

$$z_k = \sum_{i=1}^{q} g_i(\boldsymbol{\Theta})\theta_{ik}$$

è il valore della trasformata di Woodruff calcolato sull'unità *k*, dove θ_{ik} è il valore assunto dalla variabile ϑ_i sull'unità medesima. Quindi, per la stima della varianza si utilizza l'approssimazione

$$V(\tilde{Y}) \doteq V(\hat{Z}), \qquad (2.5)$$

in cui

$$\hat{Z} = \sum_{i=1}^{q} g_i(\boldsymbol{\Theta})\hat{\Theta}_i \qquad (2.6)$$

è uno stimatore corretto del totale *Z*.

Pertanto, data la variabile $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)$ = (Y, **X**, $\mathbf{T}_1$, $\mathbf{T}_2$), in cui $\mathbf{T}_1$ e $\mathbf{T}_2$ sono le variabili che hanno come totali rispettivamente $\boldsymbol{T}_1$ e $\boldsymbol{T}_2$, applicando quanto appena visto allo stimatore di regressione generalizzata, ponendo $\boldsymbol{\Theta} = (\mathrm{Y}, \boldsymbol{X}, \boldsymbol{T}_1, \boldsymbol{T}_2)$ e $\hat{\boldsymbol{\Theta}} = (\hat{Y}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{T}}_1, \hat{\boldsymbol{T}}_2)$ si ha:

$$g_1(\boldsymbol{\Theta}) = \left.\frac{\partial \hat{Y}_{GREG}}{\partial \hat{Y}}\right|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = 1,$$

$$g_2(\boldsymbol{\Theta}) = \left.\frac{\partial \hat{Y}_{GREG}}{\partial \hat{X}_j}\right|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = -\hat{B}_j\Big|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = -B_j, \qquad j = 1,\ldots,J,$$

$$g_3(\boldsymbol{\Theta}) = \left.\frac{\partial \hat{Y}_{GREG}}{\partial t_{1jj'}}\right|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = (\boldsymbol{X} - \hat{\boldsymbol{X}})'(-\hat{\boldsymbol{T}}_1^{-1}\boldsymbol{\Lambda}_{jj'}\hat{T}_1^{-1})\hat{T}_2\Big|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = 0, \qquad j \le j' = 1,\ldots,J,$$

$$g_4(\boldsymbol{\Theta}) = \left.\frac{\partial \hat{Y}_{GREG}}{\partial t_{2j}}\right|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = (\boldsymbol{X} - \hat{\boldsymbol{X}})'\hat{T}_1^{-1}\boldsymbol{\lambda}_j\Big|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = 0, \qquad j \le j' = 1,\ldots,J,$$

in cui $\Lambda_{jj'}$ è una matrice $J \times J$ con il valore 1 nella posizione (*j*, *j*') e il valore 0 altrove; λ_j è un vettore di dimensione *J* con il *j*-simo elemento pari ad 1 e tutti gli altri uguali a 0; $t_{1jj'}$ è l'elemento (*j*, *j*') della matrice $\boldsymbol{T}_1$; t_{2j} è l'elemento *j*-simo del vettore $\boldsymbol{T}_2$.

Sostituendo le derivate $g_i(\boldsymbol{\Theta})$ (*i* =1, …, 4) nella (2.6), si ottiene

$$\hat{Z} = \hat{Y} - \hat{\boldsymbol{X}}\boldsymbol{B} = \sum_{k \in s} \frac{y_k - \boldsymbol{x}_k' \boldsymbol{B}}{\pi_k} = \sum_{k \in s} \frac{z_k}{\pi_k}$$

e, dunque, si è in grado di trovare l'approssimazione di $V(\hat{Y}_{GREG})$ data dalla (2.5).

Per quanto riguarda lo stimatore della varianza di $\hat{Y}_{GREG}$, una espressione generale è data da

$$var(\hat{Y}_{GREG}) = var\left(\sum_{k \in s} \frac{y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}}{\pi_k} g_{ks} \right) = var\left(\sum_{k \in s} \frac{\hat{z}_k}{\pi_k} g_{ks} \right), \quad (2.7)$$

in cui, si introduce il termine approssimato della trasformata di Woodruff

$$\hat{z}_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}} \quad (2.8)$$

ed il fattore correttivo g_{ks} il quale permette di ottenere uno stimatore meno distorto, sotto il modello, rispetto a quello che utilizza i soli coefficienti diretti $1/\pi_k$ (Deville e Särndal, 1992).

Si può infine notare che i valori $\hat{z}_k$ coincidono con i termini e_k definiti nella (1.5). Nella precedente trattazione ci si è riferiti al caso di un modello a livello di unità elementari e di un gruppo di riferimento del modello a livello di totale popolazione. E' facile, tuttavia, adottare tale metodologia agli altri modelli descritti nel *paragrafo 1*.

3 LO STIMATORE DI REGRESSIONE GENERALIZZATA PER I DIVERSI DISEGNI DI CAMPIONAMENTO

Nel presente paragrafo sono presentate le espressioni dello stimatore di regressione $\hat{Y}_{GREG}$, e il relativo stimatore della varianza, $var\left(\hat{Y}_{GREG}\right)$, nei diversi disegni di campionamento con e senza reimmissione. Per non appesantire eccessivamente tale trattazione si esaminano direttamente le strategie campionarie che adottano un disegno stratificato, tralasciando l'analisi del caso in cui la popolazione non sia suddivisa in strati. Quest'ultimo caso, tuttavia, è facilmente riconducibile al campionamento stratificato considerando una popolazione costituita da un unico strato.

3.1 CAMPIONAMENTO DI UNITÀ ELEMENTARI CON PROBABILITÀ D'INCLUSIONE COSTANTI

Sia *U* una popolazione suddivisa in *H* strati e si indichi con:

h (*h*=1, …, *H*) l'indice del generico strato costituito da N_h unità, dove $\sum_h N_h = N$;

k (*k*=1, …, N_h) l'indice della generica unità finale di campionamento appartenente allo strato *h*;

Il parametro da stimare si può in questo caso esprimere con

$$Y = \sum_{h=1}^{H} \sum_{k=1}^{N_h} y_{hk} ,$$

dove y_{hk} rappresenta il valore assunto dalla variabile Y sull'unità elementare *k* inclusa nello strato *h*.

Si supponga di aver estratto da *U*, attraverso un disegno casuale stratificato, un campione *s*, in cui per ciascuno strato *h* la selezione delle n_h unità $(\sum_h n_h = n)$ sia stata effettuata con reimmissione e probabilità uguali. In tale contesto lo stimatore di regressione generalizzata per il totale *Y* si può scrivere come

$$\hat{Y}_{GREG} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_{hk}\ g_{hk} \qquad (3.1)$$

in cui il termine N_h/n_h rappresenta il coefficiente diretto dell'unità *k* appartenente allo strato *h* e g_{hk} è un fattore correttivo ottenuto mediante l'espressione (1.12) o alternativamente dalla (1.14), a seconda del tipo di gruppo di riferimento del modello adottato.

In base alla (2.7), si calcola la stima della varianza dello stimatore $\hat{Y}_{GREG}$ mediante l'espressione

$$\mathrm{var}\left(\hat{Y}_{GREG}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(\frac{\hat{z}_{hk} g_{hk}}{\pi_{hk}} - \tilde{\tilde{Z}}_h \right)^2 = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\ \frac{1}{n_h - 1}\ \sum_{k=1}^{n_h} \left(\hat{z}_{hk}\ g_{hk} - \tilde{\tilde{Z}}_h \right)^2 \qquad (3.2)$$

in cui $\hat{z}_{hk}$ è la trasformata di y_{hk} data dall'espressione (2.8) e dove $\tilde{\tilde{Z}}_h = {}^{1}\!/_{n_h} \sum_k^{n_h} \hat{z}_{hk} g_{hk}$.

Se la selezione delle unità nel campione avviene senza reimmissione, lo stimatore del parametro Y è dato sempre dalla (3.1), mentre la stima della varianza è calcolata tramite l'espressione

$$var\left(\hat{Y}_{GREG}\right) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \frac{1}{n_h - 1} \left(1 - \frac{n_h}{N_h}\right) \sum_{h=1}^{n_h} \left(\hat{z}_{hk}\ g_{hk} - \tilde{\tilde{Z}}_h \right)^2 = \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h}\ \frac{1}{n_h - 1}\ \sum_{h=1}^{n_h} \left(\hat{z}_{hk}\ g_{hk} - \tilde{\tilde{Z}}_h \right)^2 \qquad (3.3)$$

E' da sottolineare che nella (3.3) compare esplicitamente il termine N_h, a differenza di quanto avviene per la (3.2.) in cui non è richiesta la conoscenza diretta di N_h in quanto sostituendo π_{hk} con n_h/N_h si ottiene la prima delle (3.2) che non dipende da N_h.

In base a tale considerazione per calcolare la (3.3) bisogna, dunque, conoscere la numerosità dello strato. Tuttavia la formula (3.3) è calcolata

sostituendo N_h con la stima $\hat{N}_h$ ottenuta con i pesi diretti. Tale stima riporta esattamente al totale N_h quando tutte le unità del campione hanno risposto. In presenza del fenomeno della mancata risposta totale, nel caso in cui sono stati utilizzati come coefficienti iniziali di input i coefficienti diretti senza la correzione per mancata risposta totale, la quantità $\hat{N}_h$ sottostima il totale N_h. In presenza di mancata risposta totale, si consiglia pertanto di utilizzare i coefficienti diretti corretti per mancata risposta totale.

La (3.2) e la (3.3) rappresentano una stima corretta della varianza se $\hat{Y}_{GREG}$ è uno stimatore lineare, mentre sono consistenti per il disegno (*design consistent*) e sono approssimativamente corretti rispetto al modello di regressione sottostante se lo stimatore $\hat{Y}_{GREG}$ non è lineare (Särndal et al., 1992 pag.238; Särndal et al.,1989).

3.2 CAMPIONAMENTO A GRAPPOLI CON PROBABILITÀ D'INCLUSIONE COSTANTI

Si definisca con *U* l'universo di riferimento dei grappoli (già introdotto nel *paragrafo 1.2*) con U_I suddiviso in *H* strati e in relazione al generico strato *h* si indichi con:

i (*i*=1, …, N_h) l'indice della generico grappolo di unità elementari;

k (*k*=1, …, M_{hi}) l'indice della generica unità elementare appartenente al grappolo *i* dello strato *h*.

Inoltre, si denoti sinteticamente con (*hik*) la generica unità elementare *k* inclusa nel grappolo *i* dello strato *h*.

In questo caso il parametro si può rappresentare come

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik},$$

dove y_{hik} è il valore della variabile Y osservato sull'unità elementare (*hik*).

Sia *s* un campione di *n* grappoli ottenuto attraverso un disegno casuale stratificato, in cui per ciascuno strato si estraggono con reimmissione e probabilità uguali n_h grappoli. In questo tipo di disegno, che prevede un solo stadio di selezione ed in cui si selezionano grappoli di unità elementari, le unità primarie di campionamento coincidono con le unità finali di campionamento che sono rappresentate dai grappoli di unità elementari.

Nel campionamento a grappoli la definizione dello stimatore di regressione generalizzata varia a seconda del livello del modello utilizzato. La scelta del livello influisce sulla forma dello stimatore nella definizione del fattore correttivo.
In generale lo stimatore è espresso come

$$\hat{Y}_{GREG} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{M_{hi}} y_{hik}\, g_{hik} \,, \qquad (3.4)$$

in cui per il modello a livello di unità elementari, g_{hik} è dato dalla:

(1) (1.12), se si utilizza un unico gruppo di riferimento del modello, che coincide con l'intera popolazione;
(2) (1.14), se si utilizzano *D* (*d*=1,…,*D*) gruppi di riferimento del modello.

Per il modello a livello di cluster si ha che g_{hik} è dato dalla

(3) (1.17), se si utilizza un unico gruppo di riferimento del modello, che coincide con l'intera popolazione;
(4) (1.18), se si utilizzano *D* (*d*=1,…,*D*) gruppi di riferimento del modello.

Adattando la (2.7) a questo disegno di campionamento, la stima della varianza dello stimatore $\hat{Y}_{GREG}$, definito dalla (3.4), è calcolata con la formula seguente

$$var\left(\hat{Y}_{GREG}\right) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \; \frac{1}{n_h - 1} \; \sum_{i=1}^{n_h} \left(\tilde{Z}_{hi} - \tilde{\tilde{Z}}_h\right)^2 \,, \qquad (3.5)$$

essendo

$$\tilde{Z}_{hi} = \sum_{k=1}^{M_{hi}} \hat{z}_{hik}\; g_{hik} \qquad , \qquad \tilde{\tilde{Z}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{Z}_{hi} \;.$$

Se la selezione dei grappoli avviene senza reimmissione, lo stimatore è sempre espresso dalla (3.4), mentre la stima della sua varianza si ottiene con

$$var\left(\hat{Y}_{GREG}\right)=\sum_{h=1}^{H}\frac{N_h(N_h-n_h)}{n_h}\ \frac{1}{n_h-1}\ \sum_{i=1}^{n_h}\left(\tilde{Z}_{hi}-\tilde{\bar{Z}}_h\right)^2 . \qquad (3.6)$$

Nella (3.6) valgono le stesse considerazioni esposte in relazione alla (3.3) per quanto riguarda il termine N_h.

Le espressioni (3.5) e (3.6) rappresentano stimatori corretti (o approssimativamente corretti se la funzione è non lineare) della varianza campionaria, nel caso in cui si adatta uno stimatore $\hat{Y}_{GREG}$ espresso dalla (3.4).

3.3 CAMPIONAMENTO DI UNITÀ ELEMENTARI CON PROBABILITÀ D'INCLUSIONE VARIABILI

In presenza di un disegno con probabilità di inclusione variabili lo stimatore del totale Y si presenta come:

$$\hat{Y}_{GREG}=\sum_{h=1}^{H}\sum_{k=1}^{n_h}\frac{y_{hk}\ g_{hk}}{\pi_{hk}}, \qquad (3.7)$$

in cui si è indicato con π_{hk} la probabilità d'inclusione dell'unità *k* nello strato *h* e g_{hk} è il fattore correttivo ottenuto tramite la (1.12) o la (1.14). La (3.7) rappresenta un'espressione più generale della (3.1) ed è valida per un disegno di campionamento con o senza reimmissione.

Secondo la (2.7), lo stimatore adottato è

$$var\left(\hat{Y}_{GREG}\right)=\ \sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{k=1}^{n_h}\left(\frac{\hat{z}_{hk}\ g_{hk}}{\pi_{hk}}-\tilde{\bar{Z}}_h\right)^2 , \qquad (3.8)$$

essendo

$$\tilde{\bar{Z}}_h=\frac{1}{n_h}\sum_{k=1}^{n_h}\frac{\hat{z}_{hk}\ g_{hk}}{\pi_{hk}} .$$

Lo stimatore (3.8) risulta corretto (o approssimativamente corretto se $\hat{Y}_{GREG}$ è non lineare) nel caso in cui il campione sia stato selezionato con reimmissione, mentre risulta distorto se il campione è stato selezionato senza reimmissione, determinando delle stime approssimate per eccesso. Tuttavia, è necessario sottolineare che la distorsione è trascurabile quando il tasso di campionamento all'interno degli strati è "piccolo" (Wolter,1985).

La scelta di non utilizzare lo stimatore corretto (o approssimativamente corretto se $\hat{Y}_{GREG}$ è non lineare) della varianza quando la selezione delle unità è senza reimmissione, è dettata dalla difficoltà di calcolo delle probabilità di inclusione di secondo ordine delle unità, le quali sono necessarie per definire tale stimatore. Ulteriori considerazioni sull'uso dell'espressione (3.8) per disegni senza reimmissione sono state evidenziate nel *capitolo 3*.

3.4 CAMPIONAMENTO A GRAPPOLI CON PROBABILITÀ D'INCLUSIONE VARIABILI

Lo stimatore $\hat{Y}_{GREG}$ in tale contesto assume la forma:

$$\hat{Y}_{GREG} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{k=1}^{M_{hi}} \frac{y_{hik}\, g_{hik}}{\pi_{hik}} \,, \qquad (3.9)$$

essendo per l'unità (*hik*):

π_{hik} la probabilità d'inclusione costante per tutte le unità elementari appartenenti al grappolo *i* dello strato *h*, e pari alla probabilità di inclusione π_{hi} dello stesso grappolo *i*;

g_{hik}, il fattore correttivo che si può esprimere alternativamente con la (1.12), la (1.14), la (1.17) o la (1.18) a seconda che si usino o no i gruppi di riferimento ed a seconda del livello del modello prescelto.

Per gli analoghi motivi descritti nel caso del campionamento di unità elementari con probabilità d'inclusione variabili, si applica la stima della varianza del caso con reimmissione, anche quando si è adottato uno schema di selezione

senza reimmissione; per la stima della varianza dello stimatore (3.9), la formula impiegata è

$$var\left(\hat{Y}_{GREG}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\tilde{Z}_{hi}-\tilde{\bar{Z}}_h\right)^2 , \qquad (3.10)$$

essendo

$$\tilde{Z}_{hi}=\sum_{k=1}^{M_{hi}}\frac{1}{\pi_{hik}}\hat{z}_{hik}\ g_{hik} \qquad , \qquad \tilde{\bar{Z}}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}\tilde{Z}_{hi} \ .$$

La (3.10) risulta corretta quando la selezione avviene con reimmissione ed è distorta quando la selezione dei grappoli è senza reimmissione.

3.5 CAMPIONAMENTO A DUE O PIÙ STADI

Si consideri in una prima fase un disegno a due stadi, e sia, quindi, *U* l'universo di riferimento delle UPS suddiviso in *H* strati e in relazione al generico strato *h* si indichi con:

i (*i*=1, …, N_h) l'indice della generica UPS;

k (*k*=1, …, M_{hi}) l'indice della generica unità elementare di secondo stadio (USS) appartenente all'unità primaria *i*.

Inoltre, analogamente a quanto visto nel precedente paragrafo, si denoti sinteticamente con (*hik*) la generica USS *k* inclusa nella UPS *i* dello strato *h*.

Il parametro da stimare è, quindi, dato da

$$Y=\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}y_{hik},$$

dove y_{hik} è il valore della variabile Y osservato sull'unità elementare (*hik*).

Prendiamo in esame il caso della selezione delle UPS con probabilità variabili e siano rispettivamente: n_h il numero di UPS selezionate nello strato *h* e m_{hi} il numero delle USS selezionate nella UPS *i* dello strato *h*.

In tale contesto lo stimatore $\hat{Y}_{GREG}$, sia nel caso di selezione della UPS con reimmissione che in quello senza reimmissione, è dato dalla seguente espressione

$$\hat{Y}_{GREG} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} \frac{y_{hik}\, g_{hik}}{\pi_{hik}} \qquad (3.11)$$

dove la probabilità di inclusione π_{hik} della generica USS *(hik)* è data dal prodotto tra la probabilità di inclusione π_{hi} della UPS (*hi)* e la probabilità di inclusione condizionata $\pi_{k|hi}$ della stessa USS *(hik),* dato che al primo stadio è stata selezionata la UPS (*hi)*.

La stima della varianza di $\hat{Y}_{GREG}$ calcolata con la stessa formula, sia per la selezione con reimmissione che per quella senza reimmissione delle UPS, è data da

$$var\left(\hat{Y}_{GREG}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h}\left(\tilde{Z}_{hi} - \tilde{\bar{Z}}_h\right)^2 , \qquad (3.12)$$

essendo

$$\tilde{Z}_{hi} = \sum_{k=1}^{m_{hi}} \frac{1}{\pi_{hik}}\, \hat{z}_{hik}\; g_{hik} , \qquad \tilde{\bar{Z}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \tilde{Z}_{hi} \; .$$

In un disegno che prevede per le UPS probabilità di inclusione di primo ordine variabili, la (3.12) rappresenta uno stimatore corretto (o approssimativamente corretto se $\hat{Y}_{GREG}$ è non lineare) nel caso che il campione sia stato selezionato con reimmissione e presenta, invece, una distorsione positiva qualora la selezione delle UPS sia stata compiuta senza reimmissione.

In quest'ultimo caso l'uso della (3.12) è giustificato dalla difficoltà di calcolo delle probabilità di inclusione di secondo ordine delle UPS, richieste per definire lo stimatore corretto della varianza (o approssimativamente corretto se $\hat{Y}_{GREG}$ è non lineare).

Nel caso in cui le UPS siano estratte con probabilità di inclusione costanti, si utilizza sempre la (3.12) che è uno stimatore corretto per la selezione delle

UPS con reimmissione e distorto positivamente per la selezione delle UPS senza reimmissione.

Per disegni a tre o più stadi di campionamento non si presentano differenze sostanziali. Gli stadi di campionamento ulteriori al secondo sono integrati nella (3.11) attraverso l'inserimento di altre sommatorie per tenere conto delle unità selezionate nel campione negli stadi successivi, mentre la stima della varianza si ottiene sempre con la (3.12).

4 PRESENTAZIONE SINTETICA DEGLI ERRORI DI CAMPIONAMENTO MEDIANTE MODELLI REGRESSIVI

4.1. INTRODUZIONE

Una informazione completa sul livello di precisione dei risultati prodotti da un indagine campionaria richiederebbe la specificazione degli errori campionari di tutte le stime pubblicate. Tuttavia, le indagini su larga scala prodotte dai principali centri di diffusione statistica a livello nazionale ed internazionale sono caratterizzate da strategie campionarie complesse - basate su disegni campionari ad uno o più stadi di selezione, con stratificazione delle unità primarie selezione delle unità con probabilità variabili e senza reimmissione, da stimatori che sono funzioni non lineari dei dati campionari - e da un numero estremamente elevato di stime prodotte. Risulterebbe, quindi, oneroso e di difficile attuazione, per limiti di tempo e di costi di elaborazione, pubblicare per ciascuna stima il corrispondente errore campionario. Inoltre, le tavole di pubblicazione sarebbero appesantite e di non facile consultazione per l'utente finale.

Tali difficoltà hanno portato allo studio di alcuni metodi approssimati che agevolano notevolmente il calcolo degli errori campionari ed idonei modelli che consentono di esporre in forma concisa i suddetti errori. Tali modelli si possono suddividere in due tipi, a seconda della metodologia utilizzata: quella dei *modelli regressivi* e quella basata sull'*effetto del disegno di campionamento* (o *deft*, *design effect*) (Verma, Scott e Muircheartaigh, 1980; Verma, 1982; Wolter, 1985). La metodologia implementata è quella dei modelli regressivi ed è fondata sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento.

L'approccio utilizzato per la costruzione dei modelli regressivi è differente a seconda che le stime di interesse siano:

(i) stime di frequenze assolute o relative, riferite alle modalità di una variabile qualitativa, oppure alle classi formate in base ad una variabile quantitativa; esempi di stime di questo tipo sono:

- la stima del numero totale di individui della popolazione che risultano occupati, oppure la stima del numero totale di individui appartenenti alla classe di età [10-12) anni;
- la stima del numero totale di imprese della popolazione che producono un dato tipo di prodotto, oppure la stima del numero totale di imprese che appartengono alla classe dimensionale [1-3) addetti;

(ii) stime di totali di variabili quantitative; esempi di stime di questo tipo sono:

- il valore monetario complessivo delle spese effettuate dalle famiglie italiane nel mese di dicembre, oppure il numero totale di viaggi di lavoro effettuati dagli individui della popolazione italiana nel primo trimestre dell'anno;
- il totale degli addetti che lavorano nelle imprese italiane, oppure il totale degli investimenti effettuati da tali imprese.

Per le stime del tipo (i) è possibile utilizzare modelli regressivi che hanno un fondamento teorico, secondo cui gli errori relativi delle stime di frequenze sono espressi da una funzione decrescente al crescere dei valori delle stime stesse. Per le stime del tipo (ii), invece, il problema è piuttosto complesso, dal momento che non è stata ancora elaborata un'adeguata base teorica per l'interpolazione degli errori campionari delle stime in questione. L'approccio adottato per trattare il caso di variabili quantitative è pertanto di tipo empirico ed è fondato sull'evidenza sperimentale che l'errore assoluto di un totale è una funzione crescente del totale stesso.

Nel seguito del paragrafo verranno descritti separatamente i modelli regressivi adottati per le stime del tipo (i) ed (ii). Una trattazione approfondita degli argomenti di seguito trattati è contenuta anche nel lavoro di Russo (1987).

4.2. CARATTERISTICHE GENERALI DEL METODO

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichino rispettivamente con $V(\hat{Y}_\omega)$ e con $\sigma(\hat{Y}_\omega) = \sqrt{V(\hat{Y}_\omega)}$, la varianza e l'errore di campionamento della stima $\hat{Y}_\omega$ del generico parametro di interesse Y_ω $(\omega = 1,...,\Omega)$; si indichino, inoltre, con

$$\varepsilon^2(\hat{Y}_\omega) = \frac{V(\hat{Y}_\omega)}{Y_\omega^2} \quad , \quad \varepsilon(\hat{Y}_\omega) = \frac{\sigma(\hat{Y}_\omega)}{Y_\omega}$$

le corrispondenti quantità relative.

Denotato con $G = \left\{\hat{Y}_\omega, (\omega = 1,...,\Omega)\right\}$ l'insieme delle stime di interesse, l'ipotesi fondamentale alla base del metodo dei modelli regressivi è quella che, nell'ambito dell'insieme G, l'errore campionario relativo, $\varepsilon(\hat{Y}_\omega)$, oppure la varianza campionaria relativa, $\varepsilon^2(\hat{Y}_\omega)$, dipendono soltanto dall'ampiezza del parametro Y_ω. Ad esempio è possibile definire un legame funzionale che lega la varianza relativa $\varepsilon^2(\hat{Y})$ di una stima $\hat{Y}$, con il corrispondente valore del parametro di interesse, Y, mediante la seguente relazione funzionale:

$$\varepsilon^2\left(\hat{Y}\right) = f\left(Y, \alpha_1,...,\alpha_q, u\right) \qquad (4.1)$$

in cui $\alpha_1,...,\alpha_q$ sono dei parametri incogniti e u è un errore casuale.

In pratica la precedente relazione viene sostituita dall'analoga relazione operativa

$$\hat{\varepsilon}^2\left(\hat{Y}\right) = f\left(\hat{Y}, \alpha_1,...,\alpha_q, u\right) \qquad (4.2)$$

in cui

$$\hat{\varepsilon}^2\left(\hat{Y}\right) = \frac{\hat{V}(\hat{Y})}{\hat{Y}^2}$$

La stima dei parametri $\alpha_1,...,\alpha_q$ si ottiene adattando il modello (4.2) ad una nuvola di punti $\left(\hat{Y}_\omega, \hat{\varepsilon}^2\left(\hat{Y}_\omega\right)\right)$ formata da un sotto insieme, $G' = \left\{\hat{Y}_\omega, (\omega = 1,...,\Omega')\right\}$ di numerosità Ω' ($\Omega' \leq \Omega$), delle stime appartenenti all'insieme G e dalle corrispondenti varianze relative $\left\{\varepsilon^2\left(\hat{Y}_\omega\right), (\omega = 1,...,\Omega')\right\}$.

Si perviene, pertanto, al seguente modello stimato

$$\hat{\varepsilon}^2\left(\hat{Y}\right) = f\left(\hat{Y}, \hat{\alpha}_1,...,\hat{\alpha}_q, e\right) \qquad (4.3)$$

in cui $\hat{\alpha}_1,...,\hat{\alpha}_q$ indicano rispettivamente le stime dei parametri incogniti $\alpha_1,...,\alpha_q$ ed e rappresenta il residuo ottenuto come

$$e = \hat{\varepsilon}^2\left(\hat{Y}\right) - \hat{\hat{\varepsilon}}^2\left(\hat{Y}\right)$$

essendo $\hat{\hat{\varepsilon}}^2\left(\hat{Y}\right)$ il corrispondente valore stimato della varianza relativa della stima $\hat{Y}$, ottenuto attraverso la relazione

$$\hat{\hat{\varepsilon}}^2\left(\hat{Y}\right) = f\left(\hat{Y}, \hat{\alpha}_1,...,\hat{\alpha}_q\right)$$

Per ciascuna stima appartenente all'insieme G è possibile, quindi, determinare una stima della corrispondente varianza relativa mediante la relazione

$$\hat{\hat{\varepsilon}}^2\left(\hat{Y}_\omega\right) = f\left(\hat{Y}_\omega, \hat{\alpha}_1,...,\hat{\alpha}_q\right) \qquad (4.4)$$

A partire dalla (4.4) è possibile, poi, ottenere l'errore relativo ed assoluto, espressi rispettivamente da

$$\hat{\hat{\varepsilon}}\left(\hat{Y}_\omega\right) = \sqrt{f\left(\hat{Y}_\omega, \hat{\alpha}_1,...,\hat{\alpha}_q\right)} \qquad (4.5)$$

$$\hat{\hat{\sigma}}\left(\hat{Y}_\omega\right) = \hat{\hat{\varepsilon}}\left(\hat{Y}_\omega\right)\hat{Y}_\omega \qquad (4.6)$$

Al fine di permettere il calcolo degli errori campionari delle stime pubblicate, mediante il metodo appena descritto, nei volumi in cui vengono presentati i

risultati di un indagine campionaria viene riportata, usualmente, una tabella del seguente tipo:

Tabella 14: coefficienti stimati del modello (4.2) e grado di adattamento del modello a livello totale e per ciascun dominio di studio

	Coefficienti stimati del modello			Indice di determinazione %
Totale	$\hat{\alpha}_1$	…..	$\hat{\alpha}_q$	R^2
Dominio di studio 1	$\hat{\alpha}_{1,1}$	…..	$\hat{\alpha}_{q,1}$	R_1^2
…	…	…..	…	…
Dominio di studio D	$\hat{\alpha}_{1,D}$	…..	$\hat{\alpha}_{q,D}$	R_D^2

in cui, con riferimento a ciascun dominio di studio *d* (*d=1,…,D*) e per il totale della popolazione sono contenuti i valori dei coefficienti stimati $\hat{\alpha}_1,...,\hat{\alpha}_q$. Al fine di documentare il grado di rappresentatività degli errori campionari stimati in base al modello (4.2), in tale tabella viene riportato, con riferimento a ciascun dominio di studio *d*, il coefficiente di determinazione $R_d^{\,2}$ che rappresenta il grado di adattamento della funzione interpolata alla nuvola di punti $\left(\hat{Y}_{\omega,d}\,,\hat{\varepsilon}^2\left(\hat{Y}_{\omega,d}\right)\right)$.

Poiché per gli utenti non statistici il calcolo degli errori campionari mediante i modelli interpolati (4.5) può risultare di non facile utilizzo, si affianca generalmente alla tabella 14 una tabella che permette una valutazione più agevole degli errori campionari delle stime pubblicate, anche se conduce a risultati meno precisi. La suddetta tabella, che viene presentata con riferimento a ciascun dominio di studio, è del seguente tipo:

Tabella 15: valori interpolati degli errori relativi in corrispondenza ad alcuni valori tipici prefissati delle stime, a livello totale e per ciascun dominio di studio

Dominio di studio 1		…….		Dominio di studio D		Totale	
Livelli di stima prefissati	Errori relativi interpolati	…	…	Livelli di stima prefissati	Errori relativi interpolati	Livelli di stima prefissati	Errori relativi interpolati
$\hat{Y}^*_{1,1}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{1,1}\right)$	…	…	$\hat{Y}^*_{1,D}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{1,D}\right)$	$\hat{Y}^*_{1}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{1}\right)$
.	.	.	…	.	.	.	.
.	.	.	.	.	.	.	.
$\hat{Y}^*_{k,1}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{k,1}\right)$	…	…	$\hat{Y}^*_{k,D}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{k,D}\right)$	$\hat{Y}^*_{k}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{k}\right)$
.	.	.	…	.	.	.	.
.	.	.	…	.	.	.	.
$\hat{Y}^*_{K,1}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{K,1}\right)$	…	…	$\hat{Y}^*_{K,D}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{K,D}\right)$	$\hat{Y}^*_{K}$	$\hat{\hat{\varepsilon}}\left(\hat{Y}^*_{K}\right)$

Nella prima e nella seconda colonna della tabella 15 sono riportati rispettivamente:

- alcuni particolari livelli di stima; così, ad esempio, nel caso dell'indagine Multiscopo, per la stima di frequenze assolute riferite alle famiglie si utilizzano i seguenti livelli di stima: 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 750, 1000, 2000, 3000, 4000, 5000, 7500, 15000, 20000 e 25000 migliaia, sia con riferimento a ciascun dominio di stima considerato che con riferimento al totale popolazione. Poiché in tal caso la colonna relativa alle stime $\hat{Y}^*_{k,d}$ (*k*=1,….*K*), dove *K* è l'indice del parametro d'interesse, è sempre la stessa per tutti i domini di studio *d* ed anche per il totale popolazione, la struttura della tabella 15 sopra riportata viene leggermente modificata in

quanto la colonna relativa alle stime $\hat{Y}^*_{k,d}$ viene riportata nella tabella una sola volta per tutti i domini anziché per ciascun dominio separatamente;

- i corrispondenti valori dell'errore relativo riferiti ad un particolare dominio di studio *d* ed al totale popolazione, ottenuti attraverso il modello (4.5) ponendo rispettivamente $\hat{Y}^*_{k,d} = \hat{Y}_\omega$ (per *d=1,…,D*) e $\hat{Y}^*_k = \hat{Y}_\omega$.

Nelle tabelle 14 e 15 sopra descritte, per quanto riguarda la definizione dei valori $\hat{Y}^*_{k,d}$ (*k*=1,….*K* e *d*=1,…*D*) della tabella 15, si opera nel seguente modo:

- per ciascun dominio *d* si calcola il totale popolazione T_d , ottenuto come somma dei pesi finali (COEFFIN) delle unità elementari appartenenti al dominio stesso;
- si calcola il totale popolazione, *T* , ottenuto come somma dei pesi finali (COEFFIN) di tutte le unità elementari intervistate;
- si definiscono alcuni valori tipici prefissati di stime di frequenze percentuali P^*_k (per P^*_k =0,1; 0,5; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 15; 20; 25; 30; 35; 40; 45; 50)
- si calcolano i corrispondenti valori delle stime di frequenze assolute mediante le seguenti formule

$$\hat{Y}^*_{k,d} = P_k \, T_d \quad (k=1,\ldots,K \text{ e } d=1,\ldots,D)$$

e

$$\hat{Y}^*_k = P_k \, T \,,$$

riferite rispettivamente al generico dominio *d* ed al totale popolazione.

Il calcolo dell'errore relativo corrispondente alla generica stima $\hat{Y}_{\omega,d}$ appartenente all'insieme G_d delle stime pubblicate con riferimento al dominio *d* può essere ricavato, a partire dalla tabella 15, in base ad uno dei seguenti metodi:

(1) il primo metodo consiste nell'individuare, sulla colonna della tabella 15 riferita al dominio *d*, il livello di stima che più si avvicina alla stima di interesse $\hat{Y}_{\omega,d}$ e

nel considerare come errore relativo il valore che si trova sulla stessa riga della seconda colonna della tabella riferita a detto dominio di studio;

(2) nel secondo metodo, l'errore campionario della stima $\hat{Y}_{\omega,d}$ si ricava mediante la seguente espressione

$$\hat{\hat{\varepsilon}}(\hat{Y}_{\omega,d}) = \hat{\hat{\varepsilon}}(\hat{Y}^*_{k-1,d}) - \frac{\hat{\hat{\varepsilon}}(\hat{Y}^*_{k-1,d}) - \hat{\hat{\varepsilon}}(\hat{Y}^*_{k,d})}{\hat{Y}^*_{k-1,d} - \hat{Y}^*_{k,d}} (\hat{Y}_{\omega,d} - \hat{Y}^*_{k-1,d})$$

dove $\hat{Y}^*_{k-1,d}$ e $\hat{Y}^*_{k,d}$ sono i valori delle stime, riportati nella prima colonna della tabella 15 riferita al dominio *d*, entro i quali è compresa la stima di interesse $\hat{Y}_{d,\omega}$ ed $\hat{\varepsilon}(\hat{Y}^*_{k-1,d})$ e $\hat{\varepsilon}(\hat{Y}^*_{\mathrm{k,d}})$ sono i corrispondenti errori relativi letti sulla seconda colonna della tabella, sempre riferita al dominio *d*.

E' importante sottolineare il fatto che il metodo dei modelli regressivi richiede il calcolo degli errori relativi su un sottoinsieme di stime di dimensione molto minore rispetto a quella dell'insieme G_d e tale metodo, pertanto, costituisce una semplificazione e una riduzione dei costi notevole rispetto al criterio di specificare accanto ad ogni stima pubblicata il corrispondente errore di campionamento. Nel caso delle stime di frequenze assolute per l'adattamento del modello, si presceglie generalmente per ciascun dominio di stima e per il totale popolazione un sottoinsieme di circa 40 stime di interesse distribuito in modo da coprire uniformemente l'intero campo di variabilità delle stime oggetto di pubblicazione.

4.3. IL CASO DELLE STIME DI FREQUENZE

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichi con

$$Y = \sum_{i=1}^{N} Y_i \qquad (4.7)$$

il numero totale di unità della popolazione che possiedono una data caratteristica di interesse, in cui: Y_i è una variabile indicatrice pari ad uno se l'unità *i*-esima della popolazione presenta il carattere di interesse e zero altrimenti; *N* indica la numerosità totale della popolazione di interesse. Sia inoltre

$$\hat{Y} = \sum_{i=1}^{n} K_i Y_i \qquad (4.8)$$

una stima corretta del parametro *Y* in cui

$$W_i = \frac{1}{\pi_i}$$

è il peso diretto assegnato alla *i-esima* unità campionaria ottenuto in base al disegno campionario complesso adottato e π_i rappresenta la probabilità di inclusione nel campione dell'unità *i-esima*.

La varianza campionaria della stima $\hat{Y}$ può essere espressa dal prodotto della varianza di un campione casuale semplice di numerosità *n* per la statistica *deff* (effetto del disegno di campionamento) espresso dal quadrato del *deft*. Si ha pertanto che:

$$V\left(\hat{Y}\right) = N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n} deff \qquad (4.9)$$

essendo

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \frac{1}{N} \sum_{i=1}^{N} Y_i \right)^2 . \qquad (4.10)$$

Nel caso in esame, la precedente quantità può essere riscritta come

$$\sigma^2 = P(1-P) \qquad (4.11)$$

in cui

$$P = \frac{Y}{N}$$

In base alle espressioni (4.11) e (4.9), la varianza relativa della stima $\hat{Y}$ può, quindi, essere espressa da

$$\varepsilon^2\left(\hat{Y}\right) = \frac{V\left(\hat{Y}\right)}{Y^2} = \frac{N^2}{Y^2}\frac{N-n}{N-1}\frac{P(1-P)}{n}deff \qquad (4.12)$$

che attraverso semplici passaggi assume la forma

$$\varepsilon^2\left(\hat{Y}\right) = \frac{1}{n}\frac{N-n}{N-1}\left[\frac{N}{Y}-1\right]deff \quad . \qquad (4.13)$$

Ponendo

$$A = \frac{N}{n}\frac{(N-n)}{N-1} \qquad (4.14)$$

si ottiene infine

$$\varepsilon^2\left(\hat{Y}\right) = -\frac{A}{N}\ deff + A\ deff\ \frac{1}{Y} \qquad (4.15)$$

Sotto l'ipotesi che il *deff* sia costante (o approssimativamente tale), nell'ambito di un determinato insieme *G* di stime di frequenze assolute, è possibile formulare un modello regressivo del tipo (4.2) per stimare l'errore

campionario delle stime appartenenti a tale insieme. In base alla (4.15) si ha, quindi, che

$$\varepsilon^2\left(\hat{Y}\right) = \alpha_1 + \frac{\alpha_2}{Y} + u \qquad (4.16)$$

E' possibile ottenere un modello alternativo al precedente modificando opportunamente la (4.15). Si ottiene, infatti, mediante semplici passaggi che tale relazione può essere riscritta come

$$\varepsilon^2\left(\hat{Y}\right) = -\frac{A\,deff}{Y}\left(1-\frac{Y}{N}\right). \qquad (4.17)$$

Calcolando il logaritmo di entrambi i membri della precedente relazione si ottiene

$$\log\left(\varepsilon^2\left(\hat{Y}\right)\right) = \log\left(A\,deff\right) - \log\left(Y\right) + log\left(1-\frac{Y}{N}\right) \qquad (4.18)$$

La precedente relazione non è lineare in $\log\left(\varepsilon^2\left(\hat{Y}\right)\right)$ e $\log\left(Y\right)$ per la presenza del terzo termine a secondo membro, tuttavia per valori bassi del rapporto (*Y/N*) tale termine è trascurabile. Pertanto, sotto l'ipotesi che il *deff* sia costante nell'ambito dell'insieme di stime *G*, si ottiene il seguente modello alternativo

$$\log\left(\varepsilon^2\left(\hat{Y}\right)\right) = \alpha_1 + \alpha_2 \log\left(Y\right) + u \qquad (4.19)$$

per stimare la varianza delle stime appartenenti all'insieme *G*.

Il corrispondente modello non lineare è espresso quindi da

$$\varepsilon^2\left(\hat{Y}\right) = \tilde{\alpha}_1\ Y^{\tilde{\alpha}_2}\ \tilde{u} \qquad (4.20)$$

in cui si è posto

$$\tilde{\alpha}_1 = anti\log(\alpha_1),\ \tilde{\alpha}_2 = \alpha_2 \qquad (4.21)$$

e

$$\tilde{u} = anti\log(u).$$

Una stima dei parametri α_1 e α_2 del modello (4.19) si ottiene, mediante il metodo dei minimi quadrati (semplici o ponderati, nel caso in cui viene rilasciata l'ipotesi di omoschedasticità), adattando il modello in oggetto ad una nuvola di punti $\left(\hat{Y}, \hat{\varepsilon}^2\left(\hat{Y}\right)\right)$ formata da un sotto insieme G' di stime, appartenenti all'insieme *G*, e dalle corrispondenti varianze relative.

Si perviene, in tal modo , al seguente modello stimato

$$\log\left(\hat{\varepsilon}^2\left(\hat{Y}\right)\right) = \hat{\alpha}_1 + \hat{\alpha}_2 \log\left(\hat{Y}\right) + e \qquad (4.22)$$

in cui $\hat{\alpha}_1$ e $\hat{\alpha}_2$ indicano rispettivamente gli stimatori dei minimi quadrati dei parametri incogniti α_1 e α_2 ed e rappresenta il residuo ottenuto come

$$e = \hat{\varepsilon}^2\left(\hat{Y}\right) - \hat{\hat{\varepsilon}}^2\left(\hat{Y}\right)$$

essendo $\hat{\hat{\varepsilon}}^2\left(\hat{Y}\right)$ il corrispondente valore stimato della varianza relativa della stima $\hat{Y}$, ottenuto attraverso la relazione

$$\log\left(\hat{\varepsilon}^2\left(\hat{Y}\right)\right) = \hat{\alpha}_1 + \hat{\alpha}_2 \log\left(\hat{Y}\right) \qquad (4.23)$$

E' possibile ottenere una stima dei parametri $\tilde{\alpha}_1$ e $\tilde{\alpha}_2$ del modello non lineare (4.20) sfruttando le relazioni (4.21). Si ha pertanto

$$\hat{\tilde{\alpha}}_1 = anti\log(\hat{\alpha}_1),\ \hat{\tilde{\alpha}}_2 = \hat{\alpha}_2 . \qquad (4.24)$$

Si ritiene importante mettere in luce il fatto che gli stimatori dei minimi quadrati $\hat{\alpha}_1$ e $\hat{\alpha}_2$ sono stimatori non distorti dei rispettivi parametri α_1 e α_2 mentre gli stimatori $\hat{\tilde{\alpha}}_1$ e $\hat{\tilde{\alpha}}_2$, del corrispondente modello non lineare, non godono della proprietà di correttezza con riferimento ai parametri $\tilde{\alpha}_1$ e $\tilde{\alpha}_2$. L'applicazione del metodo dei minimi quadrati a funzioni linearizzate dei parametri viene spesso effettuata per comodità di calcolo, poiché i metodi di stima non lineare sono più complessi.

E' possibile utilizzare il modello (4.23) anche per la presentazione sintetica di stime di frequenze relative. Infatti, per ogni stima di frequenza assoluta, $\hat{Y}$, a cui corrisponde una stima della frequenza relativa $\hat{P}$, vale la ben nota relazione

$$\hat{\hat{\varepsilon}}^2\left(\hat{P}\right) = \hat{\hat{\varepsilon}}^2\left(\hat{Y}\right) .$$

In base al modello (4.23) è possibile, quindi, scrivere

$$\log\left(\hat{\hat{\varepsilon}}^2\left(\hat{P}\right)\right) = \log\left(\hat{\hat{\varepsilon}}^2\left(\hat{Y}\right)\right) = \hat{\alpha}_1 + \hat{\alpha}_2 \log\left(\hat{Y}\right) \qquad (4.25)$$

4.4. IL CASO DELLE STIME DI TOTALI DI VARIABILI QUANTITATIVE

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichi con *Y*, espresso mediante la formula (4.7), il totale della variabile quantitativa Y in cui Y_i rappresenta il valore assunto da detta variabile con riferimento alla *i*-esima unità della popolazione di interesse; sia, inoltre $\hat{Y}$, espresso mediante la formula (4.8), una stima corretta del parametro *Y*.

Nel caso in esame, a partire dalla (4.9), sfruttando le seguenti espressioni

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - \frac{Y}{N}^2\right), \qquad (4.26)$$

$$\left(\sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N} Y_i^2 + 2\sum_{i=1}^{N}\sum_{i'>i}^{N} Y_i Y_{i'} \qquad (4.27)$$

la varianza della stima $\hat{Y}$ può essere espressa come

$$V\left(\hat{Y}\right) = \left[\frac{(N-1)}{N} AY^2 - 2A\left(\sum_{i=1}^{N}\sum_{i'>i}^{N} Y_i Y_{i'}\right)\right] deff \ , \qquad (4.28)$$

in cui *A* è dato dalla (4.14); passando alla varianza relativa si ottiene quindi

$$\varepsilon^2\left(\hat{Y}\right) = \left[\frac{(N-1)}{N} A - \frac{2A}{Y^2}\left(\sum_{i=1}^{N}\sum_{i'>i}^{N} Y_i Y_{i'}\right)\right] deff \qquad (4.29)$$

Sotto l'ipotesi che il deff sia costante (o approssimativamente) nell'ambito di un determinato insieme *G* di stime, è possibile formulare un modello regressivo del tipo (4.2) per stimare l'errore campionario delle stime appartenenti a tale insieme, che, in base alla (4.29), può essere espresso da

$$\varepsilon^2\left(\hat{Y}\right) = \alpha_1 + \frac{\alpha_2}{Y^2} + u \quad . \qquad (4.30)$$

Per tenere conto della presenza del termine $\sum_{i=1}^{N}\sum_{i'>i}^{N} Y_i Y_{i'}$ a secondo membro della (4.29) è possibile introdurre nel modello l'ipotesi di eteroschedasticità; pertanto con riferimento alle stime appartenenti all'insieme *G*, tale ipotesi è espressa da

$$E\left(u_{\omega}^{2}\right)=\sigma_{\omega}^{2}=\sigma^{2} f\left(\sum_{i=1}^{N}\sum_{i'>i}^{N} Y_{\omega,i}Y_{\omega,i'}\right) \qquad (\omega=1,...,\Omega) \qquad (4.31)$$

In presenza dell'ipotesi di eteroschedasticità, una stima efficiente e corretta dei parametri α_1 e α_2 del modello (4.30) è ottenuta in base al metodo dei minimi quadrati ponderati; per l'applicazione di tale metodo, tuttavia, sarebbe necessario conoscere le varianze σ_{ω}^{2} $(\omega=1,...,\Omega)$, oppure disporre di una loro stima. La stima delle varianze σ_{ω}^{2} può comportare, tuttavia, un aumento notevole delle difficoltà di calcolo.

Per le ragioni sopra esposte si ricorre spesso a modelli empirici che mostrano un buon adattamento ai dati osservati. Un modello empirico, che usualmente conduce a buoni risultati, è il seguente

$$\sigma\left(\hat{Y}\right)=\alpha_1+\alpha_2\hat{Y}+\alpha_3\hat{Y}^{2}+u \quad . \qquad (4.32)$$

Poiché il modello (4.32) è di tipo empirico, la stima dei parametri α_1 , α_2 e α_3 deve essere ottenuta in base ad una nuvola di punti formata utilizzando tutte le stime incluse nell'insieme *G*. Ciò è differente dalla procedura adottata nel caso delle stime di frequenze, in cui i parametri del modello vengono stimati in base ad una nuvola di punti formata da un sottoinsieme G' delle stime d'interesse. Nella situazione esaminata, infatti, la procedura adottata per le stime di frequenze non garantisce il buon adattamento del modello stesso anche alle stime dell'insieme *G* che non appartengono a G' .

A partire dalla (4.32) è possibile, quindi, stimare l'errore relativo di campionamento di una generica stima appartenente all'insieme *G* mediante le seguente espressione

$$\hat{\hat{\varepsilon}}\left(\hat{Y}\right)=\hat{\alpha}_2+\frac{\hat{\alpha}_1}{\hat{Y}}+\hat{\alpha}_3\hat{Y} \qquad (4.33)$$

in cui, $\hat{\alpha}_1$, $\hat{\alpha}_2$ e $\hat{\alpha}_3$ rappresentano le stime dei corrispondenti parametri α_1 , α_2 e α_3, ottenute in base al metodo dei minimi quadrati.

Esplicitando la precedente espressione rispetto al valore della stima $\hat{Y}$ si perviene alla seguente equazione di secondo grado:

$$\hat{\alpha}_1 + [\hat{\alpha}_2 - \hat{\hat{\varepsilon}}(\hat{Y})]\hat{Y} + \hat{\alpha}_3 \hat{Y}^2 = 0 \qquad (4.34)$$

le cui radici sono espresse rispettivamente da

$$\hat{Y}_1 = \frac{-[\hat{\alpha}_2 - \hat{\hat{\varepsilon}}(\hat{Y})] - \sqrt{[\hat{\alpha}_2 - \hat{\hat{\varepsilon}}(\hat{Y})]^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2\hat{\alpha}_3} \qquad (4.35)$$

$$\hat{Y}_2 = \frac{-[\hat{\alpha}_2 - \hat{\hat{\varepsilon}}(\hat{Y})] + \sqrt{[\hat{\alpha}_2 - \hat{\hat{\varepsilon}}(\hat{Y})]^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2\hat{\alpha}_3} \ .$$

Utilizzando le precedenti formule è possibile costruire una tabella alternativa (alla tabella 15 presentata nel *paragrafo 4.2*) di presentazione sintetica degli errori di campionamento la cui struttura è mostrata nel seguente esempio.

Tabella 16: valori dei totali corrispondenti ad alcuni valori tipici prefissati degli errori relativi a livello di totale popolazione e per ciascun dominio di studio

	Valori prefissati degli errori relativi percentuali		
	ε_1^*		ε_K^*
Totale	$\hat{Y}_1^*$	…..	$\hat{Y}_K^*$
Dominio di studio 1	$\hat{Y}_{1,1}^*$	…..	$\hat{Y}_{K,1}^*$
…			
Dominio di studio D	$\hat{Y}_{1,D}^*$	…..	$\hat{Y}_{K,D}^*$

In essa vengono riportati i valori delle stime $\hat{Y}^*$ ottenuti in base alla (4.34), in relazione ad alcuni valori tipici prefissati dell'errore relativo percentuale. Definito, pertanto, con ε_k^* (k=1,…,K) il generico valore prefissato dell'errore relativo, sostituendo tale valore nella (4.35), al posto di $\hat{\hat{\varepsilon}}(\hat{Y})$, è possibile ricavare il corrispondente valore della stima $\hat{Y}_k^*$ scegliendo il valore assunto dalla corrispondente radice positiva dell'equazione (4.34) ottenuta mediante una delle (4.34).

La lettura di tale tabella indica che le stime con valori superiori a $\hat{Y}_k^*$ presentano valori dell'errore relativo inferiori a ε_k^*, mentre le stime che assumono valori inferiori a $\hat{Y}_k^*$ presentano valori dell'errore relativo superiori a ε_k^*. I valori di ε_k^* che vengono usualmente utilizzati per la costruzione della tabella sono 5, 10, 15, 20, 25, 30 e 35%.

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., Righi, P., 2002, *Genesees v1.*0, Proceedings of the Conference CompStat 2002 – Short Communications and Posters, Berlin August 24-th to August 28th 2002 (disponibile su CD-ROM)

Pagliuca, D. (a cura di)*,* 2004a, *Genesees v.3.0.*, *Funzione Riponderazione,* Manuale utente ed aspetti metodologici, Tecniche e Strumenti, ISTAT, n. 2.

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.