LO STIMATORE DI PONDERAZIONE VINCOLATA IN PRESENZA DI INFORMAZIONI AUSILIARIE CAMPIONARIE

Claudio Ceccarelli

Istituto Nazionale di Statistica
Direzione Centrale Indagini
sulle Condizioni e Qualità della Vita
claudio.ceccarelli@istat.it

Giovanni Maria Giorgi

Sapienza Università di Roma,
Dip.to di Statistica, Probabilità
e Statistiche Applicate
giovanni.giorgi@uniroma1.it

Alessio Guandalini

Sapienza Università di Roma,
Dip.to di Statistica, Probabilità
e Statistiche Applicate
alessio.guandalini@gmail.it

Premessa

Le principali indagini svolte dagli istituti nazionali di statistica impiegano stimatori che utilizzano variabili ausiliarie altamente correlate con le variabili di interesse per migliorare l'accuratezza, la comparabilità e la coerenza delle stime. Tra gli stimatori che ricorrono a questo tipo di informazioni, detti indiretti, vi sono lo stimatore di regressione generalizzato e lo stimatore di ponderazione vincolata (anche detto stimatore calibrato). Quest'ultimo, messo a punto da Deville e Särndal nel 1992, viene comunemente utilizzato nelle principali indagini campionarie su larga scala per le sue particolari proprietà e per la sua duttilità di impiego.

La definizione originale dello stimatore di ponderazione vincolata consente l'utilizzo di variabili ausiliarie, dette vincoli, provenienti da fonte amministrativa e quindi non affette da errore di tipo campionario. Generalmente, i vincoli utilizzati nel caso di indagini su famiglie e individui riguardano totali di variabili demografiche come popolazione per età, sesso e dominio territoriale note da fonte anagrafica; tuttavia, l'esigenza di produrre stime coerenti con indagini campionarie diverse svolte quasi simultaneamente, o con stime ottenute dalla stessa indagine in occasioni precedenti, sostenuta anche da importanti Enti (ad esempio Eurostat), ha fatto sì che nell'insieme di variabili ausiliarie fossero inserite anche delle stime affette da errore campionario. L'introduzione di variabili affette da errore campionario nell'insieme di vincoli dello stimatore calibrato ha posto l'interrogativo su come questo potesse ripercuotersi sulla stima e sull'errore della variabile di interesse.

L'obiettivo di questo lavoro è, quindi, quello di mettere a punto uno stimatore della varianza per disegni complessi in grado di misurare l'aggiunta di errore commesso con lo stimatore di ponderazione vincolata che utilizza anche variabili stimate da altre indagini campionarie e valutare l'impatto che queste hanno sia in termini di qualità che di efficienza sulle stime prodotte.

L'attenzione è stata focalizzata sulla stima del totale in quanto gli stimatori di numerosi parametri possono essere espressi come combinazione lineare di totali e la loro varianza campionaria può essere calcolata in funzione delle stime delle varianze e covarianze campionarie degli stimatori di totali.

Il lavoro è composto essenzialmente da due parti: nella prima, sono trattate le caratteristiche dello stimatore di ponderazione vincolata ed è illustrata la metodologia sviluppata per calcolare la varianza quando questo utilizza informazioni ausiliarie campionarie (con particolare attenzione per disegni campionari complessi e per indagini dipendenti); nella seconda, invece, sono presentate le espressioni derivanti dalle caratteristiche dei disegni di campionamento delle due principali indagini campionarie socio-economiche condotte dall'Istat su famiglie e individui, in modo da valutare

l'effettivo impatto che il ricorso a vincoli campionari ha sulla qualità e l'efficienza delle stime prodotte.

1 Lo stimatore di ponderazione vincolata

Lo stimatore di ponderazione vincolata appartiene alla classe di stimatori indiretti che costituiscono un metodo di stima valido per garantire il miglioramento dell'accuratezza, correttezza e comparabilità delle stime. Questa classe di stimatori consente di raggiungere questi obiettivi facendo ricorso ad informazioni ausiliarie campionarie altamente correlate con la variabile di interesse e per le quali sono noti i totali riferiti alla popolazione di riferimento o a particolari partizioni di questa.

Nelle indagini su larga scala, su ciascuna unità i del campione s $(i = 1, \dots, k, \dots, n)$ vengono rilevate, oltre alla variabile di interesse Y, una serie di variabili ausiliarie $(j = 1, \dots, J)$ sintetizzate dal vettore $\boldsymbol{x} = \left(x_{1k}, \dots, x_{jk}, \dots, x_{Jk}\right)'$, così per ciascuna unità campionaria si ha un vettore $(y_k, \boldsymbol{x}_k)$ composto da $J + 1$ osservazioni.

Lo stimatore di ponderazione vincolata, messo a punto da Deville e Särndal, nella sua definzione originaria, permette di sfruttare l'informazione dei totali delle variabili ausiliarie riferita alla popolazione oggetto d'indagine per produrre stime coerenti (vincolate) per individui e famiglie (Deville *et al.*, 1992). Si assume che il vettore dei totali delle variabili ausiliarie corrispondenti alle J variabili ausiliarie rilevate, $\boldsymbol{X} = \sum_{k \in U} \boldsymbol{x}_k = \left(\sum_{k \in U} \boldsymbol{x}_{1k}, \dots, \sum_{k \in U} \boldsymbol{x}_{jk}, \dots, \sum_{k \in U} \boldsymbol{x}_{Jk}\right)$, sia noto e sia riferito all'intera popolazione oggetto d'indagine o eventualmente a particolari sottopopolazioni.

L'idea alla base di questo stimatore è che "…*weights that perform well for the auxiliary variable also should perform well for study variable*" (Deville *et al.*, 1992, p. 376). Si utilizza, infatti, la calibrazione sui totali noti nella popolazione delle variabili ausialiarie $\boldsymbol{X}$ per correggere i pesi base assegnati alle unità dallo stimatore di Horvitz-Thompson (HT) in base al disegno campionario prescelto. Lo stimatore HT garantisce la correttezza delle stime attribuendo a ciascuna unità campione un peso pari al reciproco della probabilità di inclusione del primo livello $(w_k = d_k = \pi_k^{-1})$ (Cicchitelli *et al.*, 1992, p.102).

Lo stimatore di ponderazione vincolata, della forma:

$$\hat{Y} = \sum_{k=1}^{n} w_k y_k \,, \tag{1.1}$$

assegna lo stesso coefficiente di riporto all'universo w_k $(w_k = d_k \gamma_k)$ a tutti gli individui dello stesso nucleo familiare (Lemaître *et al.*, 1987, p. 199) in modo che questo sia, in media, per una data funzione di distanza, il più vicino possibile al peso base d_k assegnato secondo il disegno di

campionamento (Deville *et al.*, 1993, p. 1013), rispettando però un sistema di vincoli sintetizzato dalla relazione:

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{X}\,. \tag{1.2}$$

La calibrazione sui pesi delle J variabili ausiliarie X si basa sull'assunto che maggiore è la correlazione con il parametro che si intende stimare, maggiore è l'accuratezza della stima che si ottiene. L'intento di modificare il meno possibile i pesi base, $d_k = \pi_k^{-1}$, invece, è giustificato dal fatto di voler mantenere l'importante proprietà dello stimatore HT di condurre a stime non distorte di tali pesi, che equivale a minimizzare, per ogni campione s, la distanza tra i due pesi

$$\left\{\sum_{k \in s} G(w_k, d_k)\right\}, \tag{1.3}$$

dove $G(\cdot)$ rappresenta la funzione di distanza prescelta per la misurazione.

La soluzione di un sistema di minimo vincolato del tipo:

$$\begin{cases} min\left\{\sum_{k \in s} G(w_k, d_k)\right\} \\ \sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{X} \end{cases} \tag{1.4}$$

porta all'individuazione di un insieme di pesi w_k che continueranno a dare stime non distorte.

La soluzione del sistema di minimo vincolato esiste se la funzione di distanza $G(w_k, d_k)$ è strettamente crescente e continua, quindi, se esiste la funzione inversa $g_k^{-1}(\cdot)$ tale che $w_k = g_k^{-1}\big(g(w_k, d_k)\big)$. Alla soluzione numerica, data dal vettore $\boldsymbol{w} = (w_1, \dots, w_k, \dots, w_n)'$, si giunge dopo aver impostato la *funzione di Lagrange* e determinato il vettore $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k, \dots, \lambda_n)$ in modo iterativo mediante il *metodo di Newton*. Il processo iterativo si interrompe quando la maggior differenza relativa sui valori λ tra un'iterazione e la precedente è minore di un valore piccolo a piacere compreso nell'intervallo $(0,1)$, oppure nel caso in cui si superano il numero di iterazioni massime consentite (generalmente fissato a 50), oltre il quale si giudica che l'algoritmo non converge (Pagliuca, 2005a, pp. 145-149).

Determinati i valori dei λ è possibile calcolare il fattore di correzione del peso base per ciascun individuo che è uguale a:

$$\gamma_k = F_k\big(\boldsymbol{x}_k'\lambda\big) = \frac{1}{d_k} g_k^{-1}\big(\boldsymbol{x}_k'\,\lambda\big)\,, \tag{1.5}$$

cioè ad una funzione della variabile $u_k = \big(\boldsymbol{x}_k'\,\lambda\big)$ combinazione lineare del vettore di variabili ausiliarie $\boldsymbol{x}_k$ e dei J valori incogniti del vettore $\boldsymbol{\lambda}$. L'espressione funzionale dello stimatore di ponderazione vincolata dell'ammontare del carattere Y, dunque, è:

$$\hat{Y}_{PV} = \sum_{k\epsilon s} w_k \gamma_k y_k, \qquad (1.6)$$

con γ_k uguale alla (1.5).

1.1 La funzione di distanza

Un ruolo importante nella definizione dello stimatore calibrato è giocato dalla funzione utilizzata per misurare la distanza tra i pesi finali e i pesi base. La scelta di questa non solo influenza l'intervallo di variazione dei pesi finali, ma anche l'esistenza di una soluzione nel caso in cui il sistema di vincoli sia congruente.

Nella letteratura specialistica sull'argomento sono note diverse funzioni per l'aggiustamento dei pesi base che vengono utilizzate per lo stimatore di ponderazione vincolata. Una rassegna di queste, e delle relative proprietà, è fornita da Deville (1992, pp. 377-378) e da Singh (1996, pp. 108-114). Le più utilizzate sono, ad esempio:

a. Euclidea o Lineare $\dfrac{(w_k - d_k)^2}{2d_k}$

b. Lineare Troncata $\begin{cases} \dfrac{(w_k - d_k)^2}{d_k} & \text{se } L < \dfrac{w_k}{d_k} < U \\ \infty & \text{altrimenti} \end{cases}$

c. Logaritmica $w_k ln\left(\dfrac{w_k}{d_k}\right) - w_k + d_k$

d. Logaritmica Troncata o Logit $\left(\dfrac{w_k}{d_k} - L\right) ln\left(\dfrac{\frac{w_k}{d_k} - L}{1 - L}\right) + \left(U - \dfrac{w_k}{d_k}\right) ln\left(\dfrac{U - \frac{w_k}{d_k}}{U - 1}\right)$

e. Chi-Quadrato (modificato) $\dfrac{1}{2w_k}\left(\dfrac{w_k}{d_k} - 1\right)^2$

f. Minima Entropia $-d_k ln\left(\dfrac{w_k}{d_k}\right) + w_k - d_k$

g. Hellinger $2d_k\left(\sqrt{\dfrac{w_k}{d_k}} - 1\right)^2$.

Tutte le funzioni illustrate soddisfano le proprietà di regolarità descritte nel precedente paragrafo ma, soprattutto, sono tutte asintoticamente equivalenti allo stimatore di regressione generalizzato. In tutte, infatti, il modello per il fattore di aggiustamento è lineare in $\boldsymbol{x}$:

$$F_k\left(\boldsymbol{x}_k'\lambda\right) = \left(1 + a\boldsymbol{x}_k'\lambda\right)^{1/a}, \qquad (1.7)$$

con a che assume valori diversi per le sette funzioni di distanza illustrate sopra.

Alcune di queste funzioni, però, possono portare alla determinazione di pesi finali negativi o in alcuni casi troppo elevati, che diventano difficilmente gestibili nelle indagini su larga scala. Una funzione di distanza che ovvia a questo problema, e proprio per questo è quella correntemente utilizzata nelle indagini su larga scala, si ottiene ponendo $a = (U - L/(U - 1)(1 - L)$. Tale funzione, nota come logaritmica troncata, consente di ottenere pesi compresi in un intervallo "accettabile" attraverso la definizione *a priori* del valore minimo (L) e massimo (U) che possono assumere i coefficienti di correzione γ_k.

1.2 Convergenza asintotica dello stimatore calibrato

Come detto sopra, per ciascuna funzione di distanza, F_k e il fattore di correzione sono sempre lineari in $\boldsymbol{x}$. Questo ha consentito a Deville e Särndal di dimostrare che lo stimatore che deriva dalle diverse funzioni di distanza è asintoticamente equivalente allo stimatore di regressione generalizzato $(GREG)$ (Deville *et al.*, 1992, pp.379-380).

Lo stimatore $GREG$, infatti, può essere visto, asintoticamente, come un caso particolare in cui lo stimatore calibrato è generato da $F_k = 1 + \boldsymbol{x}_k^{'}\lambda$, ovvero il caso specifico dello stimatore PV che deriva dalla funzione di distanza lineare. Questo risultato può essere esteso a tutte le funzioni di distanza illustrate precedentemente, e quindi anche alla funzione logaritmica troncata, poiché per queste il modello del fattore di aggiustamento è lineare rispetto al vettore di variabili ausiliarie.

Nel caso di campioni di grandi dimensioni, dunque, è possibile ricorrere allo stimatore $GREG$ per studiare le proprietà dello stimatore PV, soprattutto nel caso della varianza, che altrimenti non sarebbe possibile studiare analiticamente a causa della complessità che il calcolo della probabilità di inclusione del secondo ordine può assumere.

1.3 La varianza dello stimatore di ponderazione vincolata

Prima di andare ad illustrare l'espressione funzionale della varianza dello stimatore PV (equivalente asintoticamente a quella dello stimatore $GREG$) è opportuno, poiché tornerà utile in seguito, ricordare l'espressione e le caratteristiche principali dello stimatore $GREG$.

Nella scrittura più usuale:

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\boldsymbol{X} - \widehat{\boldsymbol{X}}_{HT})'\widehat{\mathbf{B}}\,, \qquad (1.8)$$

lo stimatore $GREG$ è ottenuto come somma dello stimatore HT della variabile di interesse e di un termine di aggiustamento dato dalle differenze tra i totali noti e le corrispondenti stime campionarie HT delle variabili ausiliarie ponderate con i rispettivi coefficienti di regressione stimati attraverso l'espressione (Pagliuca, 2005b, pp. 163-166):

$$\widehat{\mathbf{B}} = \left(\hat{\beta}_1, \dots, \hat{\beta}_j, \dots, \hat{\beta}_J\right)' = \left(\sum_{k\epsilon s} \frac{\boldsymbol{x}_k \boldsymbol{x}'_k}{\pi_k c_k}\right)^{-1} \sum_{i\epsilon U} \frac{\boldsymbol{x}_k y_k}{\pi_k c_k} \ . \tag{1.9}$$

L'idea alla base dello stimatore $GREG$ è che, considerando variabili ausiliarie correlate con la variabile di interesse, è possibile ottenere uno stimatore più efficiente dello stimatore HT (Bethlehem *et al.*, 1987, p. 143). Lo stimatore $GREG$, dunque, sfrutta le informazioni ausiliarie attraverso la definizione di un modello che spiega la nuvola di punti individuata dall'insieme $\{(y_k, \boldsymbol{x}_k): k = 1, \dots, N\}$ (Wright, 1983, p. 879; Estevao, 1995, p. 184).

La varianza asintotica dello stimatore $GREG$, seguendo Deville *et al.*(1992, pp. 379-380), può essere scritta nella forma:

$$Var\left(\hat{Y}_{PV}\right) = \sum_{k\in U} \sum_{l\neq k} \Delta_{kl}(e_k)(e_l), \tag{1.10}$$

e una stima corretta di tale quantità è:

$$var\left(\hat{Y}_{GREG}\right) = \sum_{k\in s} \sum_{l\neq k} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_{ks} \frac{e_k}{\pi_k}\right)\left(\gamma_{ls} \frac{e_l}{\pi_l}\right) \tag{1.11}$$

dove $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Per quanto detto prima, dunque, quest'ultima espressione può essere utilizzata per stimare la varianza dello stimatore di ponderazione vincolata e per studiarne le sue proprietà.

2 La varianza dello stimatore *PV* con l'aggiunta di vincoli campionari

Lo stimatore calibrato garantisce, per definizione, la coerenza con una fonte esterna al riparo da errori campionari, quindi amministrativa o censuaria, in quanto la stima dei totali è vincolata a totali noti con certezza, $\boldsymbol{X}$. Nulla vieta, però, di utilizzare stime che provengano da altre indagini campionarie (Deville, 1999, p. 207).

Vincolando i totali al vettore $\widehat{\widehat{\boldsymbol{X}}}$ $\left(\widehat{\widehat{\boldsymbol{X}}} = \widehat{\widehat{X}}_1, \dots, \widehat{\widehat{X}}_j, \dots, \widehat{\widehat{X}}_J\right)$ di stime delle medesime quantità calcolate con dati rilevati in un'altra indagine si ottiene il sistema di minimo vincolato per lo stimatore di ponderazione che utilizza variabili ausiliarie campionarie, di seguito indicato con $\widehat{PV}$:

$$\begin{cases} min\left\{\sum_{k\in s} G(w_k, d_k)\right\} \\ \sum_{k\in s} w_k \boldsymbol{x}_k = \widehat{\widehat{\boldsymbol{X}}} \end{cases} \tag{2.1}$$

dove nella relazione di coerenza non solo il vettore che si stima $\left(\widehat{\boldsymbol{X}} = w_k \boldsymbol{x}_k\right)$ è affetto da errore campionario, ma anche il vettore dei totali noti, $\widehat{\widehat{\boldsymbol{X}}}$, a cui si vincola tale stima.

Questo non comporta variazioni nell'espressione funzionale della stima del totale, ma rende necessario studiare ulteriormente le proprietà, ovvero la varianza, dello stimatore calibrato così definito, poiché non è detto che le stime ottenute attraverso la calibrazione con altre stime siano più efficienti. Le proprietà dello stimatore $\widehat{PV}$ sono studiate facendo ricorso alla convergenza asintotica allo stimatore $GREG$ seguendo un approccio *condizionato* e *non condizionato* (Ballin *et al.*, 2000, pp. 48-51).

Nell'approccio non condizionato si assume che il vettore delle stime $\widehat{\widehat{\boldsymbol{X}}}$ sia una costante e, quindi, la varianza dello stimatore $\widehat{PV}$ è ricondotta a quella dello stimatore $GREG$, presentato nella (1.10). Questa ipotesi fa sì che lo stimatore così costituito sia vincolato a totali che potrebbero risultare distorti, quindi, le proprietà dello stimatore $\widehat{PV}$ del totale non possono essere studiate valutando solo la sua varianza ma è necessario valutare anche la sua distorsione:

$$bias\left(\hat{Y}_{\widehat{PV}}\right) = \mathbf{B}'\left(\widehat{\widehat{\mathbf{X}}} - \mathbf{X}\right) \tag{2.2}$$

e, di conseguenza, il suo MSE, dato dalla somma della varianza e dal quadrato della sua distorsione, è:

$$MSE\left(\hat{Y}_{\widehat{PV}}\right) = \sum_{k\in s}\sum_{l\neq k}\frac{\Delta_{kl}}{\pi_{kl}}\left(\gamma_{ks}\frac{e_k}{\pi_k}\right)\left(\gamma_{ls}\frac{e_l}{\pi_l}\right) + \left[\mathbf{B}'\left(\widehat{\widehat{\mathbf{X}}} - \mathbf{X}\right)\right]^2$$

Nell'ottica condizionata, invece, si valutano le proprietà dello stimatore tenendo conto anche della variabilità delle stime desunte dalla fonte esterna campionaria a cui sono vincolate le stime dei totali per l'indagine in questione. Adattando lo stimatore di regressione della forma della (1.8), nel caso di variabili ausiliare desunte da altre indagini e quindi affette anch'esse da errore campionario, questo assume la forma:

$$\hat{Y}_{\widehat{GREG}} = \hat{Y}_{HT} + \mathbf{B}'\left(\widehat{\widehat{\mathbf{X}}} - \widehat{\mathbf{X}}_{\mathrm{HT}}\right). \tag{2.3}$$

Questa quantità (Ballin *et al.*, 2000, p. 48) può essere scissa nella somma di due elementi:

$$A_1 = \hat{Y}_{HT} + \mathbf{B}'\left(\mathbf{X} - \widehat{\mathbf{X}}_{\mathrm{HT}}\right)$$

$$A_2 = \mathbf{B}'\left(\widehat{\widehat{\mathbf{X}}} - \mathbf{X}\right).$$

La varianza di A_1, essendo questa quantità uguale allo stimatore $GREG$ visto nella (1.8), coincide con la varianza espressa dalla (1.10) e una sua stima, anche questa già vista in precedenza, è data da:

$$var(A_1) = \sum_{k\in s}\sum_{l\neq k}\frac{\Delta_{kl}}{\pi_{kl}}\left(\gamma_{ks}\frac{e_k}{\pi_k}\right)\left(\gamma_{ls}\frac{e_l}{\pi_l}\right).$$

Per A_2, che nella (2.2) è stata indicato anche come $bias\left(\hat{Y}_{\widehat{PV}}\right)$, la stima della varianza è pari a:

$$var(A_2) = \sum_{j=1}^{J} \mathrm{B}_j^2 var\left(\hat{\hat{X}}_j\right) + \sum_{j=1}^{J} \sum_{j' \neq j} \mathrm{B}_j \mathrm{B}_{j'} cov\left(\hat{\hat{X}}_j, \hat{\hat{X}}_{j'}\right) \quad (2.4)$$

e non dipende dal campione s, ma dal campione dell'indagine da cui vengono desunte le stime di totali utilizzati come vincoli.

È necessario operare un'ulteriore distinzione per differenziare il caso in cui le quantità A_1 e A_2 sono indipendenti o dipendenti. Sono indipendenti se l'indagine di riferimento non è legata da alcuna relazione di dipendenza con l'indagine da cui si prendono i totali noti delle variabili ausiliarie, o meglio i campioni delle due indagini sono distinti. Sono dipendenti, invece, quando le due indagini, o le due occasioni di indagine, hanno in comune, in maniera non casuale, delle unità campionarie.

Nel caso in cui le due indagine sono indipendenti tra loro, studiato da Ballin *et al*. (2000), anche le quantità A_1 e A_2 sono indipendenti e una stima asintoticamente corretta della varianza è:

$$var\left(\hat{Y}_{\widehat{PV}}\right) \cong var\left(\hat{Y}_{\widehat{GREG}}\right) = var(A_1) + var(A_2) =$$

$$= \sum_{k \in s} \sum_{l \neq k} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_{ks} \frac{e_k}{\pi_k}\right)\left(\gamma_{ls} \frac{e_l}{\pi_l}\right) + \sum_{j=1}^{J} \mathrm{B}_j^2 var\left(\hat{\hat{X}}_j\right) + \sum_{j=1}^{J} \sum_{j' \neq j} \mathrm{B}_j \mathrm{B}_{j'} cov\left(\hat{\hat{X}}_j, \hat{\hat{X}}_{j'}\right), \quad (2.5)$$

dove, il primo termine misura l'errore nella stime della variabile Y per l'indagine in questione e gli altri due termini l'errore aggiuntivo che si commette vincolando la stima dei totali delle variabili ausiliarie ad un'altra indagine affetta a sua volta da errori campionari.

Per studiare il caso in cui, invece, i campioni delle due indagini coincidono totalmente o in parte, poiché A_1 e A_2 non possono essere considerate indipendenti, è necessario tener conto, oltre che delle varianze delle due componenti prese singolarmente, anche della "contemporaneità" della variazione di A_1 e A_2, $\left(cov(A_1, A_2)\right)$, e di un fattore di correzione, f, che sconta, per tutti i membri dell'espressione in cui entra in gioco A_2 (la componente dell'altra indagine), le unità campionarie presenti in entrambe le rilevazioni, ovvero tutti quegli individui per cui agisce il vincolo $\hat{\hat{X}}_j$.

Quindi:

$$var\left(\hat{Y}_{\widehat{PV}}\right) = \sum_{k \in s} \sum_{l \neq k} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_{ks} \frac{e_k}{\pi_k}\right)\left(\gamma_{ls} \frac{e_l}{\pi_l}\right) +$$

$$+ \sum_{j=1}^{J} f_j \left(\mathrm{B}_j^2 var\left(\hat{\hat{X}}_j\right) + \sum_{j' \neq j} \mathrm{B}_j \mathrm{B}_{j'} cov\left(\hat{\hat{X}}_j, \hat{\hat{X}}_{j'}\right) + 2cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) \right) \quad (2.6)$$

dove, per quanto detto prima, $cov(A_1, A_2)$ è stata esplicitata attraverso l'espressione $cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right)$.

L'efficienza delle stime dello stimatore calibrato con vincoli campionari può essere studiata semplicemente attraverso l'espressione (2.5) o (2.6), in quanto modificando il meno possibile i pesi base per ciascuna unità, $\hat{Y}_{\widehat{PV}}$ mantiene, praticamente inalterate, alcune proprietà dello stimatore *HT*, tra cui la non distorsione delle stime.

2.1 La variabilità congiunta di $\widehat{\boldsymbol{Y}}_{\boldsymbol{GREG}}$ e $\boldsymbol{bias}\left(\widehat{\boldsymbol{Y}}_{\widehat{\boldsymbol{PV}}}\right)$

La quantità $cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right)$, vista nella (2.6), rappresenta la variabilità congiunta dovuta al fatto che le stime calcolate e i vincoli campionari utilizzati derivano da campioni che hanno in comune una quota o la totalità degli individui. L'introduzione di questa quantità rispetto alle formulazioni in Ballin *et al*. (2000), dunque, consente di stimare la varianza dello stimatore $\widehat{PV}$ anche nel caso in cui i vincoli campionari sono desunti da indagini dipendenti, ovvero indagini che utilizzano come vincoli le stime ottenute intervistando una quota delle unità già presenti nel campione (come si vedrà in seguito per il caso della Rilevazione sulle Forze di Lavoro).

La determinazione dell'espressione analitica di questa quantità, riportata in Appendice, pur riconducendosi al caso dello stimatore $\widehat{PV}$ che utilizza una sola variabile campionaria da fonte esterna, è risultata piuttosto complessa e può essere utilizzata a seconda delle assunzioni che si fanno sulla stima del vettore dei coefficienti di regressione e sulla correttezza della stima del totale della variabile ausiliaria proveniente da una fonte esterna campionaria.

Risultato 1. Se il coefficiente di regressione B è una variabile casuale e la stima del totale utilizzata come vincolo è corretta:

$$cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) = Y + B + X - 2BX - 2YB - 2YX + 4YBX$$

se, invece, è distorta $\left(E\left[\hat{\hat{X}}\right] = X + b\right)$ allora:

$$cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) = Y + B + X + b - 2BX - Bb - 2YB - 2YX - 2Yb + 4YBX + 2YBb$$

Risultato 2. Se il coefficiente che lega la variabile ausiliaria X alla variabile di interesse Y può essere considerata una costante per la stabilità che mantiene nel tempo, nel caso in cui $\hat{\hat{X}}$ è corretto:

$$cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) = Y\ ;$$

nel caso in cui, invece, è non corretto:

$$cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) = Y + Bb - 2YBb\ .$$

3 La varianza dello stimatore *PV* con vincoli campionari nei diversi disegni di campionamento

In questo paragrafo si vogliono illustrare le espressioni dello stimatore calibrato del totale con vincoli campionari, $\hat{Y}_{\widehat{PV}}$, e il relativo stimatore della varianza, $var(\hat{Y}_{\widehat{PV}})$, nei diversi disegni campionari, ma soprattutto per quei disegni che effettivamente vengono utilizzati nelle indagini su larga scala (Pagliuca, 2005b, pp. 203-211; Zannella, 1989, pp. 45-94).

Le espressioni illustrate di seguito fanno riferimento al caso più semplice in cui le indagini sono indipendenti e i vincoli campionari agiscono per tutti gli individui del campione. Per adattarle al caso di indagini dipendenti è sufficiente aggiungere l'elemento di variabilità congiunta mentre, per il caso in cui i vincoli agiscono solo per una quota del campione, basta scontare la componente di variabilità esterna per la quota f.

Come già evidenziato nel §1.2, la convergenza asintotica dello stimatore $\widehat{PV}$ allo stimatore $\widehat{GREG}$ e il fatto che si fa riferimento a indagini con campioni di elevate dimensioni consentono, nella seguente trattazione, di utilizzare in maniera equivalente la notazione $\hat{Y}_{\widehat{GREG}}$ e $\hat{Y}_{\widehat{PV}}$.

La stima della varianza del totale nei vari disegni campionari è data dalla somma delle componenti di variabilità dell'indagine in questione (quindi la varianza dello stimatore $GREG$, o PV, senza vincoli campionari) e dalla variabilità introdotta dalle variabili ausiliarie campionarie esterne $\left(var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right)\right)$ e quindi dovuta all'altra indagine:

$$var(\hat{Y}_{\widehat{PV}}) = var(\hat{Y}_{\widehat{GREG}}) = var(\hat{Y}_{GREG}) + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right). \quad (3.1)$$

Nei disegni complessi sarà possibile disaggregare questa quantità fino al dominio minimo pianificato (ad esempio provinciale o regionale) solo laddove vi è perfetta corrispondenza tra il criterio di stratificazione e la collocazione dei comuni negli strati delle due diverse indagini.

3.1 Campione casuale semplice

Nel caso di campionamento casuale semplice, viene estratto da una popolazione composta da N individui un campione formato da n unità con o senza reimmissione. Nell'estrazione con ripetizione le unità hanno sempre la stessa probabilità, $1/N$, in ciascuna delle n estrazioni, nel caso senza ripetizione, invece alla prima estrazione avranno probabilità $1/N$, alla seconda $1/N-1$, e così via. Considerando che il parametro da stimare è:

$$Y = \sum_{k=1}^{n} y_k$$

per entrambi i casi lo stimatore è:

$$\hat{Y}_{\widehat{GREG}} = \frac{N}{n}\sum_{k=1}^{n} y_k \gamma_{ks}, \tag{3.2}$$

dove N/n è il peso diretto associato ad ogni unità e γ_{ks} il suo fattore correttivo.

La stima della varianza è data dall'espressione:

$$var(\hat{Y}_{\widehat{GREG}}) = \frac{N^2}{n}\frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{z}_k \gamma_{ks} - \hat{\bar{Z}}\right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right). \tag{3.3}$$

nel caso con reimmissione e

$$var(\hat{Y}_{\widehat{GREG}}) = \frac{N}{n}\frac{N-n}{n-1}\sum_{k=1}^{n}\left(\hat{z}_k \gamma_{ks} - \hat{\bar{Z}}\right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right) \tag{3.4}$$

nel caso senza reimmissione con $\hat{\bar{Z}} = \frac{1}{n}\sum_{k=1}^{n} \hat{z}_k \gamma_{ks}$.

3.2 Campione stratificato

Per programmare il campione in corrispondenza di determinate subpopolazioni (domini di studio) al fine di contenere l'errore atteso delle stime, si ricorre alla suddivisione della popolazione in strati non sovrapposti. La popolazione U viene, dunque, suddivisa in H strati in base a delle variabili di stratificazione in modo da avere $\sum_{h=1}^{H} N_h = N$ e, da ciascuno strato, si estrae, in maniera indipendente rispetto a quello che succede per gli altri strati, un campione di numerosità n_h, tale che $\sum_{h=1}^{H} n_h = n$.

Il parametro da stimare è dato dalla somma dei valori assunti dalla variabili Y per ciascun individuo di ciascuno strato:

$$Y = \sum_{h=1}^{H}\sum_{k=1}^{n_h} y_{hk}.$$

Le unità elementari, però, possono essere selezionate con o senza ripetizione. Nel caso di selezione con reimmissione delle unità all'interno dello strato lo stimatore di regressione asintoticamente equivalente allo stimatore ponderato con vincoli campionari è:

$$\hat{Y}_{\widehat{GREG}} = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k=1}^{n_h} y_{hk}\gamma_{hk}, \tag{3.5}$$

dove, in maniera speculare alla (3.2), N_h/n_h è il peso diretto dell'unità k-ma appartenente allo strato h e γ_{hk} è il suo fattore correttivo. La stima della varianza all'interno di ciascuno strato è:

$$var\left(\hat{Y}^{h}_{\widehat{GREG}}\right) = \frac{N_h^2}{n_h}\frac{1}{n_h - 1}\sum_{k=1}^{n_h}\left(\hat{z}_{hk}\gamma_{hk} - \hat{\bar{Z}}_h\right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right), \tag{3.6}$$

con $\hat{\bar{Z}}_h = \frac{1}{n_h}\sum_{h=1}^{H}\hat{z}_{hk}\gamma_{hk}$.

Per il caso in cui le unità vengono selezionate senza reimmissione lo stimatore del parametro è quello visto nella (3.5), mentre lo stimatore della varianza all'interno dello strato h è:

$$var\left(\hat{Y}^{h}_{\widehat{GREG}}\right) = \frac{N_h}{n_h}\frac{N_h - n_h}{n_h - 1}\sum_{k=1}^{n_h}\left(\hat{z}_{hk}\gamma_{hk} - \hat{\bar{Z}}_h\right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right) \tag{3.7}$$

In entrambi i casi, poiché gli stimatori $\hat{Y}^{h}_{\widehat{GREG}}$ $(h = 1, \dots, H)$ sono indipendenti, la stima della varianza nella popolazione è data dalla somma delle stime delle varianze degli strati calcolate con le rispettive espressioni più la stima della varianza della di $bias\left(\hat{Y}_{\widehat{PV}}\right)$:

$$var\left(\hat{Y}_{\widehat{GREG}}\right) = \sum_{h=1}^{H} var\left(\hat{Y}^{h}_{\widehat{GREG}}\right) + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right). \tag{3.8}$$

3.3 Campione a grappoli

Nel campionamento a grappoli, anche detto a uno stadio, dalla popolazione U, si estrae un campione s di n grappoli. Da ogni grappolo h selezionato $(h = 1, \dots, n)$ si rilevano le informazioni su tutte le m_{hi} unità che lo compongono. Il parametro da stimare è dato da:

$$Y = \sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} y_{hik}.$$

I grappoli, però, possono avere probabilità di inclusione costanti o variabili e possono essere estratti con o senza ripetizione. Si hanno, quindi, quattro casi che portano alla definizione di altrettante espressioni funzionali per la stima di Y e della sua varianza:

a_1. probabilità di inclusione costante e estrazione con reimmissione;

a_2. probabilità di inclusione costante e estrazione senza reimmissione;

b_1. probabilità di inclusione variabile e estrazione con reimmissione;

b_2. Probabilità di inclusione variabile e estrazione senza reimmissione

In generale, nel caso di estrazione con reimmissione dei grappoli che hanno probabilità di inclusione costanti, lo stimatore è:

$$\hat{Y}_{\widehat{GREG}} = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\sum_{k=1}^{M_{hi}} y_{hik}\,\gamma_{hik}\ . \tag{3.9}$$

La stima della varianza per questo disegno di campionamento è:

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

essendo $\hat{Z}_{hi} = \sum_{k=1}^{M_{hi}} \hat{z}_{hik}\, \gamma_{hik}$ e $\hat{\bar{Z}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{Z}_{hi}$.

Se la selezione dei grappoli che hanno probabilità di inclusione costanti avviene senza reimmissione $\hat{Y}_{GREG}$ è espresso dalla (3.9), mentre la stima della varianza di tale stimatore è:

$$var(\hat{Y}_{\widehat{GREG}}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Z}_{hi} - \hat{\bar{Z}}_h \right)^2. \quad (3.11)$$

Sia questa che l'espressione della varianza nel caso di grappoli selezionati con reimmissione rappresentano degli stimatori corretti (o approssimativamente corretti se la funzione non è lineare) della varianza campionaria.

Nel caso in cui grappoli hanno probabilità di inclusione variabile lo stimatore di regressione, sia nel caso con che senza reimmissione, è:

$$\hat{Y}_{\widehat{GREG}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{k=1}^{M_{hi}} \frac{y_{hik}\, \gamma_{hik}}{\pi_{hik}}\,, \quad (3.12)$$

poiché per ciascuna unità appartenente al grappolo i dello strato h la probabilità di inclusione è pari a quella del grappolo che la contiene, π_{hi}.

La stima della varianza nel caso con reimmissione è uguale a:

$$var\left(\hat{Y}^h_{\widehat{GREG}}\right) = \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(\frac{\hat{z}_{hi}\gamma_{hi}}{p_{hi}} - \hat{\bar{Z}}_h \right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right) \right), \quad (3.13)$$

mentre, nel caso senza reimmissione è:

$$var\left(\hat{Y}^h_{\widehat{GREG}}\right) = \frac{N_h^2}{n_h} \frac{N_h - n_h}{n_{hi} - 1} \sum_{k=1}^{n_{hi}} \left(\hat{z}_{hi}\gamma_{hi} - \hat{\bar{Z}}_h \right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right) \right). \quad (3.14)$$

3.4 Campione a due stadi

Il campionamento a due stadi consente di estendere la rilevazione ad un numero più elevato di grappoli, avviene, infatti, oltre alla selezione di questi, anche la selezione con un disegno casuale semplice senza ripetizione delle unità elementari (USS) che costituiscono i grappoli (UPS). Come prima per la selezione degli n grappoli in ciascuno strato h si hanno quattro possibili scenari.

Le USS indipendentemente dal disegno campionario dei grappoli hanno probabilità costanti e vengono selezione sempre con un campionamento casuale semplice senza reimmissione, lo stimatore del totale relativamente all'i-mo grappolo è, dunque:

$$\hat{Y}_{\widehat{GREG}}^{hi} = \sum_{k=1}^{m_{hi}} \frac{y_{hik}}{\pi_{hik}} \gamma_{hik} \tag{3.15}$$

con π_{hik} probabilità di inclusione della generica unità appartenente al grappolo i-mo dello strato h uguale a M_{hi}/m_{hi}. La stima della varianza all'interno dell'i-ma PSU è della forma già vista nella (3.7), quindi:

$$var\left(\hat{Y}_{\widehat{GREG}}^{hi}\right) = \frac{M_{hi}^2}{m_{hi}} \frac{M_{hi} - m_{hi}}{m_{hi} - 1} \sum_{k=1}^{m_{hi}} \left(\hat{z}_{hik}\gamma_{hik} - \hat{\bar{Z}}_{hi}\right)^2 + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right), \tag{3.16}$$

con $\hat{\bar{Z}}_{hi} = \frac{1}{m_{hi}} \sum_{k=1}^{m_{hi}} \hat{z}_{hik}\gamma_{hik}$.

La stima del totale del parametro per il generico strato h dipende dal campionamento delle UPS all'interno degli strati. Nel caso di selezione con ripetizione di grappoli con probabilità variabili è:

$$\hat{Y}_{\widehat{GREG}}^{h} = \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} \frac{M_{hi}}{m_{hi} p_{hi} n_h} y_{hik}\gamma_{hik} \tag{3.17}$$

poichè tiene conto della probabilità di inclusione della k-ma unità, M_{hi}/m_{hi}, e del il peso attribuito al generico grappolo, $1/n_h p_{hi}$. Quando i grappoli, invece, vengono selezionati senza reimmissione lo stimatore del totale dello strato è:

$$\hat{Y}_{\widehat{GREG}}^{h} = \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{\pi_{hi}} \gamma_{hi}, \tag{3.18}$$

dove π_{hi} è la probabilità di inclusione del primo ordine dell'i-ma UPS dello strato h, ma può anche essere scritto come:

$$\hat{Y}_{\widehat{GREG}}^{h} = \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} y_{hik} \left(\frac{M_{hi}}{M_h} n_h \gamma_{hi}\right) \left(\frac{M_{hi}}{m_{hi}} \gamma_{hik}\right), \tag{3.19}$$

cioè come somma dei valori della variabile Y registrati in ciascuna unità di ciascun grappolo moltiplicati con i pesi finali dovuti rispettivamente al primo e al secondo stadio di campionamento.

La stima della varianza dello strato h, dunque, potrà essere scomposta, nella parte di variabilità dovuta al primo e al secondo stadio di campionamento, più, ovviamente, la variabilità dovuta alle variabili ausiliarie prese da un'altra indagine campionaria:

$$var\left(\hat{Y}_{\widehat{GREG}}^{h}\right) = var_1\left(\hat{Y}_{\widehat{GREG}}^{h}\right) + var_2\left(\hat{Y}_{\widehat{GREG}}^{h}\right) + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right), \tag{3.20}$$

dove $var_2\left(\hat{Y}_{\widehat{GREG}}^{h}\right)$, la variabilità dovuta all'interno del grappolo i-mo (3.16), e $var_1\left(\hat{Y}_{\widehat{GREG}}^{h}\right)$, la variabilità tra i grappoli, rispettivamente uguale alla (3.13) per il caso con ripetizione e alla (3.14) per quello senza ripetizione.

La stima del totale nella popolazione è data dalla somma dell'ammontare totale della variabile in ciascuno strato:

$$\hat{Y}_{\widehat{GREG}} = \sum_{h=1}^{H} \hat{Y}^{h}_{\widehat{GREG}}. \tag{3.21}$$

e la stima della varianza relativa al totale della popolazione è data dalla somma della varianza negli H strati indipendenti:

$$var\left(\hat{Y}_{\widehat{GREG}}\right) = \sum_{h=1}^{H} \left(var_1\left(\hat{Y}^{h}_{\widehat{GREG}}\right) + var_2\left(\hat{Y}^{h}_{\widehat{GREG}}\right)\right) + var\left(bias\left(\hat{Y}_{\widehat{PV}}\right)\right). \tag{3.22}$$

Per adattare le formule viste finora anche nel caso in cui UPS appartenenti allo stesso strato vengono estratte con un disegno campionario autoponderante basta sostituire per il caso con reimmissione $p_{hi} = 1/N_h$ e $w_{hi} = \frac{N_h}{n_h}\frac{M_{hi}}{m_{hi}}$ e per il caso senza reimmissione. $\pi_{hi} = n_h \frac{M_{hi}}{m_{hi}}$ e $w_{hik} = \frac{M_h}{n_h m_{hi}}$.

3.5 Campione a più stadi

Il campionamento a più stadi è un estensione del caso precedente nel quale all'interno di ciascun grappolo vengono selezionati con un campionamento casuale semplice altri grappoli, dai quali si possono o selezionare le unità elementari (campionamento a tre stadi) o selezionare altri grappoli per ottenere campionamenti a un numero superiore di stadi.

Il totale della variabile Y nella popolazione nella quale si svolge un campionamento a n stadi sarà data da:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{i^2=1}^{n_h^2} \cdots \sum_{i^j}^{n_h^j} \cdots \sum_{i^{n-1}}^{n_h^{n-1}} \sum_{k=1}^{m_{hijk}} y_{hii^2 \ldots i^j \ldots i^{n-1}k}$$

Per lo stadio n-mo, le unità elementari, contenute all'interno del grappolo dello stadio $n-1$, hanno sempre probabilità costanti e sono selezionate con un disegno casuale semplice senza reimmissione, quindi la stima e la varianza del totale è uguale rispettivamente alle espressioni (3.15) e (3.16). Per la stima dell'ammontare totale della variabile e la stima della varianza allo stadio $n-1$ la situazione è quella vista nel campionamento a due stadi in cui dalla UPS selezionata si estraggono m_{hi} USS, quindi si adottano le formule viste nel paragrafo precedente.

Per le stime dallo stadio $n-2$ al secondo stadio la situazione è più complicata dal punto di vista formale, poiché i grappoli sono a loro volta selezionati con un disegno casuale semplice da delle unità di livello superiore. Quindi se indichiamo con Ui^jS le unità dello stadio j-mo dal quale si

estraggono le unità dello stadio $j+1$ e che sono a loro volta estratte dallo strato $j-1$, si avrà che per ogni stadio si dovrà considerare, per la stima del totale, il peso finale dell'unità $hii^2 \dots i^{j-1}i^j i^{j+1} \dots k$ che tiene conto del campionamento di tutti gli stadi a livello inferiore e per la stima della varianza dell'errore campionario commesso ad ogni stadio.

Arrivati al primo stadio la situazione si riconduce a quella già vista nel campione a 2 stadi e sarà, quindi, sufficiente sommare i valori registrati in ciascuna delle UPS dello strato h per avere la stima del totale della variabile oggetto di studio nello strato e sommare la varianza delle unità di primo stadio, che sarà formata da $n-1$ componenti che tengono conto dell'errore campionario commesso in ogni stadio di campionamento, per ottenere la stima della varianza.

Alla stima del totale e della varianza della variabile Y, infine, si arriva sommando rispettivamente tutti i totali degli strati e le varianze misurate negli strati. A quest'ultima bisogna sempre aggiungere l'espressione che tiene conto dell'errore imputabile all'uso del vettore di totali noti da un'altra indagine campionaria.

4 L'applicazione

In Istat sono condotte due indagini su larga scala che ricorrono all'utilizzo di stimatori calibrati, definiti dalla funzione logaritmica troncata (cfr. §1.1), che vincolano le loro stime anche a totali noti campionari. In particolare, l'Indagine italiana sui redditi e condizioni di vita (It-Silc) utilizza come vincoli campionari le stime sull'occupazione e sull'istruzione prodotte dalla Rilevazione sulle Forze di Lavoro (RFL), mentre la RFL utilizza dei vincoli campionari, detti longitudinali, sulla condizione di occupazione derivanti da precedenti occasioni della rilevazione stessa. Queste rappresentano, dunque, due casi reali di quanto illustrato nel §2: indagini che utilizzano stimatori calibrati con vincoli campionari derivanti da un'indagine indipendente (It-Silc ⇒ RFL) e dipendente (RFL ⇒ RFL).

Nei prossimi paragrafi, dunque, si illustrano, prima, le principali caratteristiche dei rispettivi disegni campionari di It-Silc e della RFL[1] e, successivamente, l'adattamento a questi delle formule viste nel §3.

Nell'applicazione, infine, vengono calcolati e confrontati tra loro i valori delle stime e degli errori dei principali parametri di It-Silc e della RFL ottenuti con lo stimatore di ponderazione vincolata senza vincoli campionari e lo stimatore con vincoli campionari che, però, non tiene conto dell'aumento di variabilità che questi comportano nella valutazione dell'efficienza delle stime (approccio non condizionato)

[1] Per un quadro più completo delle due indagini si rimanda rispettivamente a Ceccarelli (2008) e Gazzelloni (2006).

Il calcolo delle stime e degli errori campionari è stato svolto in ambiente SAS con l'ausilio del software GENESEES (GENEralised software for Sampling Estimates and Error in Surveys) sviluppato in Istat proprio con queste finalità. L'utilizzo del software s'è reso opportuno in due momenti, nel calcolo dei pesi finali, attraverso la funzione di "riponderazione" e nel calcolo delle stime e degli errori con la relativa funzione.

La metodologia relativa allo stimatore calibrato è alla base del software GENESEES. Per applicare le espressioni sviluppate nei §2 e §3, però, si è dovuto fare ricorso alle proprietà asintotiche dello stimatore PV che, per campioni numerosi, come quello di It-Silc (20.928 famiglie e 52.433 individui) e della RFL (20.928 e 49.114), converge con lo stimatore di regressione generalizzato. GENESEES non determina i parametri dello stimatore $GREG$ e pertanto, è stato necessario costruire, sempre in ambiente SAS un modello di regressione per ciascuna stima e dominio di interesse. Le varie componenti utili ai fini della valutazione della qualità e dell'efficienza dello stimatore PV con totali noti campionari sono state, infine, aggregate con una procedura automatica ad hoc su un foglio di calcolo Excel.

4.1 L'indagine italiana sui redditi e condizioni di vita

L'indagine It-Silc fornisce dati trasversali su reddito, povertà, esclusione sociale e condizioni di vita e dati longitudinali su reddito, lavoro e indicatori non-monetari di esclusione sociale.

La popolazione di riferimento è formata da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente all'estero, che hanno compiuto 15 anni d'età nell'anno precedente a quello dell'intervista.

La presenza di una componente trasversale e una longitudinale ha guidato la scelta verso un disegno campionario in grado di produrre stime corrette ed efficienti in entrambi i casi: un panel della durata di quatto anni composto da quattro campioni indipendenti che ruotano annualmente.

Il disegno, prevede che il campione annuale (trasversale) sia composto in ciascun anno dai quattro campioni longitudinali che si trovano rispettivamente alla prima, seconda, terza e quarta occasione di indagine (*wave*).

Ciascuno dei quattro campioni longitudinali è formato, alla prima *wave*, da circa 8.000 famiglie. Dalla seconda *wave* in poi, il campione teorico equivale ai rispondenti della *wave* precedente e, ogni anno, le famiglie che hanno concluso la quarta *wave* escono dall'indagine e vengono sostituite da un nuovo campione di famiglie alla prima *wave*. Le famiglie di ciascun campione sono selezionate attraverso un disegno a due stadi (comuni-famiglie) con la stratificazione dei comuni a livello regionale (Ar, Nar di primo livello e Nar di secondo livello) in base alla dimensione demografica.

I principali parametri prodotti in It-Silc sono la percentuale di individui poveri e il reddito medio annuo familiare. In questo contesto, per testare la varianza dello stimatore $\widehat{PV}$ del totale, si è focalizzata l'attenzione sulla stima del reddito e in particolare su quello totale familiare netto, individuale netto, individuale da lavoro, autonomo, dipendente e da pensione.

Eurostat, per armonizzare l'indagine a livello Europeo, ha suggerito di utilizzare, per le stime dei parametri, come vincoli la popolazione residente per particolari partizioni territoriali, sesso e età e il numero di famiglie residenti. L'Istat ha seguito in parte questa direttiva ampliando, però, il numero di vincoli. Attualmente, infatti, utilizza un sistema di 163 vincoli, di cui 141 noti da fonti amministrative (variabili demografiche desunte dal Censimento (2001), Bilancio demografico, Bilancio demografico sugli stranieri) e 22 provenienti dalla RFL del quarto trimestre dell'anno di riferimento del reddito. I vincoli campionari utilizzati sono relativi alla condizione lavorativa e al livello di istruzione in modo da vincolare il reddito alle stime di variabili altamente correlate provenienti dalla fonte ufficiale del mercato del lavoro (Ceccarelli *et al*., 2007).

Il ricorso a vincoli sull'occupazione e sul livello di istruzione proveniente dalla RFL, quindi alle stime di un'indagine costruita in maniera indipendente, o ancora meglio con un campione indipendente, colloca il caso nella situazione in cui l'errore della stima di interesse è indipendente da quello del vincolo campionario. È dunque sufficiente applicare l'espressione (2.5) per stimare la varianza campionaria dell'ammontare totale dei principali parametri.

4.2 I risultati dell'applicazione sui dati It-Silc

L'applicazione sull'indagine It-Silc riguarda i dati dell'Italia relativi all'anno 2008 e si è basata su un campione di 20.928 famiglie e 52.433 individui. L'attenzione si è concentrata sia su stime a livello familiare, come il reddito familiare totale netto (fy_{tot}), che su stime a livello individuale, come il reddito individuale netto (y_{ind}), il reddito da lavoro (y_{lav}), suddiviso in reddito da lavoro dipendente (y_{dip}) e autonomo (y_{aut}), e il reddito da pensione (y_{pen}).

L'ammontare totale e l'errore, per ciascuna delle variabili considerate, sono stati stimati ricorrendo allo stimatore calibrato che utilizza come vincoli variabili demografiche provenienti da fonte anagrafica e variabili sulla condizione occupazionale e sul livello di istruzione della RFL relativa al IV trimestre del 2007.

Nella Tabella 1, sono riportate le stime calcolate ricorrendo allo stimatore di ponderazione vincolata che non utilizza i vincoli campionari e quello che li utilizza, seguendo l'approccio non condizionato e condizionato.

Tabella 1 - Stima del totale delle varie tipologie di reddito ed errori campionari relativi percentuali (CV%) con lo stimatore di ponderazione vincolata senza vincoli campionari e con vincoli campionari (Approccio non condizionato e Approccio condizionato) – Italia, 2008

Parametri di interesse		NO RCFL		RCFL		
					Approccio non condizionato	Approccio condizionato
		STIMA	CV%	STIMA	CV%	CV%
REDDITO FAMILIARE						
fy_{tot}	totale netto	728.666.713.229	0,575	725.497.084.329	0,478	0,482
REDDITO INDIVIDUALE						
y_{ind}	netto	723.096.449.584	0,557	719.916.413.899	0,444	0,449
y_{lav}	da lavoro	476.049.206.657	0,789	471.503.146.949	0,615	0,620
y_{aut}	autonomo	135.119.209.848	2,399	134.471.727.459	1,851	1,862
y_{dip}	dipendente	340.929.996.809	0,780	337.031.419.491	0,479	0,502
y_{pen}	da pensione	198.309.898.768	0,684	199.553.632.004	0,633	0,639

Sebbene i valori della stima rappresentino delle cifre di difficile interpretazione, è possibile osservare le variazioni che si ottengono introducendo vincoli campionari. La differenza tra le stime ottenute con e senza vincoli campionari può essere interpretata come una misura della distorsione della stima senza vincoli sul livello di istruzione, ma soprattutto sulla condizione occupazionale. Questa distorsione è dovuta in particolar modo alla mancanza di rappresentatività nel campione di alcune categorie occupazionali particolarmente "ostili" alle rilevazioni sul reddito, come quella degli autonomi, o della sovra-rappresentazione degli inattivi, maggiormente reperibili al momento della rilevazione.

Con riferimento proprio alla stima e all'errore del reddito degli autonomi, si può osservare un importate risultato. A fronte di una variazione della stima esigua, utilizzando variabili ausiliarie affette da errore, ma altamente correlate con la variabile reddito, si ha una drastica riduzione dell'errore relativo percentuale commesso che passa da 2,399% a 1,862%. Il guadagno in termini di efficienza dovuto ad una più corretta specificazione del modello di regressione, grazie all'inserimento dei vincoli RFL, è tale da compensare ampiamente l'aumento dell'errore dovuto alla natura campionaria dei vincoli (1.851% vs 1,862%).

4.3 La Rilevazione sulle Forze di Lavoro

La Rilevazione sulle Forze di Lavoro è una rilevazione continua effettuata in tutte le settimane dell'anno e rappresenta la fonte ufficiale del mercato del lavoro in Italia.

La popolazione di interesse è costituita da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente all'estero, mentre sono esclusi i membri permanenti delle convivenze (ospizi, istituti religiosi, caserme, ecc.).

Il disegno campionario per ciascuna rilevazione trimestrale è di tipo complesso a due stadi (comuni-famiglie) con stratificazione delle unità di primo stadio a livello provinciale in base all'ampiezza demografica e prevede uno schema di rotazione del campione del tipo $(2_T, 2_T, 2_T)$. Ciascuna famiglia viene intervistata una sola volta in una specifica settimana del trimestre (sempre la stessa per tutti i trimestri in cui la famiglia entra nel campione). Per produrre stime mensili, inoltre, l'Istat, in accordo con Eurostat, ha predisposto anche un disegno di campionamento nel tempo, in modo da garantire la rappresentatività anche dei campioni mensili.

La rotazione trimestrale dei campioni fa sì che questi, ma anche i campioni mensili, siano parzialmente sovrapposti (50% a un trimestre di distanza, 25% a tre trimestri di distanza, 50% a quattro trimestri e 25% a cinque trimestri).

Il principale obiettivo della RFL è la produzione delle stime ufficiali degli occupati e delle persone in cerca di occupazione e degli inattivi. Questi tre aggregati di interesse rappresentano una partizione della popolazione in età lavorativa (15 anni e oltre) in tre gruppi esaustivi e mutuamente esclusivi. Queste stime vengono pubblicate con cadenza trimestrale e con dettaglio regionale, ma, per i principali parametri vengono prodotte anche delle stime annuali con dettaglio provinciale e delle stime mensili a livello nazionale.

Eurostat ha chiesto ai vari istituti nazionali di statistica di rendere coerenti le stime mensili con quelle trimestrali. L'Istat ha realizzato una strategia per la costruzione dei pesi che prevede, da un lato, vincoli campionari anche nel sistema di calibrazione delle stime mensili e, dall'altro, un secondo passo di calibrazione che vincola le tre stime mensili al dato trimestrale.

Per produrre le stime mensili su occupati, disoccupati e non forze lavoro si ricorre, dunque, a un sistema di 360 vincoli, di cui 294 demografici, noti con certezza da fonte anagrafica, e 66 desunti da occasioni precedenti della stessa RFL, di cui 33 dalla rilevazione di tre mesi prima e 33 da quella di dodici mesi prima.

A causa del particolare disegno campionario della RFL, che prevede una sovrapposizione del campione a un trimestre e ad un anno (quattro trimestri) di distanza del 50% del campione teorico, il caso in questione si colloca nella situazione di indagini dipendenti. Per stimare la varianza dell'ammontare totale di occupati, disoccupati e non forze lavoro, dunque, è necessario ricorrere,

nell'applicazione, all'espressione vista nella (2.6) ed esplicitare le ipotesi relative al vettore $\widehat{\boldsymbol{B}}$ dei coefficienti di regressione per determinare la quantità $cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right)$.

Le X_j campionarie e la variabile di interesse sono variabili casuali dipendenti dal campione estratto e assumono valori diversi nelle varie rilevazioni. Nonostante questo è possibile dimostrare empiricamente, grazie alla lunga serie storica dei dati sulle Forze Lavoro, che la relazione che le lega è stabile nel tempo, quindi è lecito ipotizzare, almeno nel caso in esame, che il vettore dei coefficienti di regressione, $\widehat{\boldsymbol{B}}$ sia una costante. Essendo, inoltre, le stime $\hat{\hat{X}}_j$ anche corrette, si può considerare, dal *Risultato 2* del §2.1, $cov\left(\hat{Y}_{GREG}, bias\left(\hat{Y}_{\widehat{PV}}\right)\right) = Y$.

La stima della varianza dello stimatore $\widehat{PV}$ del totale si ottiene, dunque, introducendo questo risultato nella (2.6) da cui si ha:

$$var\left(\hat{Y}_{\widehat{PV}}\right) = \sum_{k\in s}\sum_{l\neq k}\frac{\Delta_{kl}}{\pi_{kl}}\left(\gamma_{ks}\frac{e_k}{\pi_k}\right)\left(\gamma_{ls}\frac{e_l}{\pi_l}\right) + \sum_{j=1}^{J} f_j\left(\mathrm{B}_j^2 var\left(\hat{\hat{X}}_j\right) + \sum_{j'\neq j}\mathrm{B}_j\mathrm{B}_{j'} cov\left(\hat{\hat{X}}_j, \hat{\hat{X}}_{j'}\right) + 2\hat{Y}\right).$$

4.4 I risultati dell'applicazione su i dati della RFL

Per l'applicazione sui dati della RFL si è voluto testare lo stimatore calibrato con vincoli provenienti da fonti campionarie sugli stessi aggregati per i quali vengono pubblicate le stime mensili, cioè, ammontare mensile del numero di occupati, di disoccupati e di inattivi a livello Italia distintamente per i due sessi.

A questo scopo sono stati utilizzati i dati del terzo mese del terzo trimestre 2009 (settembre 2009), relativi a un campione di 20.928 famiglie e 49.114 individui. I vincoli campionari, invece, riguardano le stime dei vari aggregati di tre e di dodici mesi prima (giugno 2009 e settembre 2008) e sono stati introdotti con l'obiettivo di rendere più stabili le stime sul mercato del lavoro. L'obiettivo, infatti, è quello di correggere i campioni delle varie rilevazioni in moda da renderli, per quanto possibile, più simili tra loro e controllare gli aspetti di variabilità esogeni riducendo le differenze stagionali e congiunturali alle sole variazioni della condizione occupazionale.

I valori delle stime e degli errori relativi (CV) ottenuti con un sistema di pesi determinati con e senza vincoli campionari sono riportati in Tabella 2.

L'introduzione dei vincoli campionari comporta una variazione leggera nell'ammontare dei vari aggregati: variazioni che risultano più esigue se messe a confronto con quelle registrate nell'applicazione su It-Silc a causa del diverso motivo che ha spinto l'introduzione di questa tipologia di informazioni ausiliarie nel sistema di vincoli.

Tabella 2 -Stima di Occupati, Disoccupati e Non Forze Lavoro ed errori campionarirelativi percentuali (CV%) con lo stimatore di ponderazione vincolata senza vincoli campionari e con vincoli campionari (Approccio non condizionato e Approccio condizionato) – Settembre 2009, Italia per sesso

	SENZA VINCOLI CAMPIONARI		CON VINCOLI CAMPIONARI		
				Approccio non condizionato	Approccio condizionato
	STIMA	CV%	STIMA	CV%	CV%
			ITALIA		
OCC	22.786.251	0,440	22.886.373	0,331	0,333
DIS	2.021.889	2,678	2.031.044	2,260	2,262
NFL	34.982.481	0,275	34.873.584	0,216	0,221
			MASCHI		
OCC	13.599.617	0,491	13.647.567	0,385	0,387
DIS	1.092.265	3,508	1.093.438	3,094	3,097
NFL	14.371.974	0,435	14.323.053	0,341	0,345
			FEMMINE		
OCC	9.186.634	0,828	9.238.806	0,610	0,612
DIS	929.624	3,926	937.606	3,259	3,262
NFL	20.610.507	0,360	20.550.531	0,275	0,281

In tutti i casi, a livello nazionale, ma anche per i due sessi, si ha un aumento dell'ammontare di occupati e disoccupati, con una conseguente diminuzione degli inattivi. L'introduzione dei vincoli longitudinali consente di stabilizzare le stime garantendo la stessa condizione di tre mesi prima e un anno prima a tutti coloro che non hanno effettivamente cambiato la propria condizione nel tempo.

L'introduzione di questi vincoli comporta un miglioramento nella *performance* dello stimatore testimoniato dalla diminuzione dell'errore campionario relativo percentuale. Il guadagno in termini di efficienza è maggiore per i disoccupati, sia a livello Italia che distintamente per maschi e femmine, e raggiunge anche 7 decimi di punto percentuale (nel caso delle femmine). Questo guadagno di efficienza, dovuto ad una più corretta specificazione del modello di regressione, grazie all'inserimento dei vincoli campionari longitudinali, è tale da compensare ampiamente l'aumento dell'errore dovuto alla natura campionaria dei vincoli, mai superiore allo 0.005%.

Rispetto al caso It-Silc, però, la diminuzione della varianza è miniore sempre a causa del diverso fine e, quindi, del diverso impatto che i vincoli campionari hanno sulla determinazione del sistema di pesi finale.

5 Conclusioni

L’espressione funzionale determinata per calcolare la varianza dello stimatore di ponderazione vincolata in presenza di informazioni ausiliarie campionarie consente di misurare l’errore aggiuntivo commesso nello stimare un parametro ricorrendo a informazioni affette anch’esse da errore.

L’impatto che l’introduzione dei vincoli campionari ha sulle stime dei parametri e sulle stime degli errori campionari è diverso a seconda della finalità per cui questi vengono inseriti e a seconda della relazione che lega l’indagine per cui queste devono essere effettuate e quella, o quelle, da cui vengono desunti i totali noti.

L’applicazione svolta sui dati di It-Silc del 2008 e su quelli della RFL del settembre 2009 rappresentano due esempi agli antipodi che mostrano come il valore delle stime e la loro efficienza, introducendo i vincoli campionari, variano in maniera diversa a seconda del motivo per cui si ricorre alle informazioni ausiliarie campionarie.

In It-Silc, infatti, l’introduzione di informazioni ausiliarie campionarie ha un impatto considerevole sia sul valore dell’ammontare totale del reddito, che diminuisce, sia sulla stima dell’errore (diminuisce anch’essa, poiché condizione occupazionale e titolo di studio sono variabili altamente correlate con il reddito). I vincoli hanno la funzione di rendere coerenti le stime con la principale fonte di informazione sul mercato del lavoro ma, soprattutto, di riequilibrare un campione che potrebbe risultare particolarmente distorto a causa di un meccanismo che genera la mancata risposta sensibilmente lontano dalla condizione di *missing at random*. Questo aspetto diventa maggiormente evidente quando si considerano i redditi da lavoro autonomo (e quindi il numero degli autonomi) e la tipologia.

Nella RFL, invece, la differenza tra il valore e l’errore commesso per occupati, disoccupati e non forze lavoro, ottenute con lo stimatore con e senza vincoli campionari, sono più esigue, in quanto, la funzione svolta dai vincoli campionari è, soprattutto, quella di rendere più stabili le stime, oltre che vincolarle a variabili per cui è dimostrato essere molto elevata la correlazione.

Le differenze dei valori ottenuti nel calcolo degli errori tra l’approccio non condizionato e condizionato, per entrambe le rilevazioni, sono piuttosto piccole (l’errore relativo dell’approccio condizionato aumenta leggermente). Nel complesso, quindi, il ricorso a stimatori di ponderazione vincolata che utilizzano tra l’insieme di vincoli anche informazioni ausiliarie campionarie, migliora la qualità della stima e comporta anche un miglioramento notevole rispetto al caso in cui il sistema non contempla questi vincoli, a patto che le variabili ausiliarie campionarie siano altamente correlate con la variabile di interesse.

Appendice

Dimostrazione $cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right)$ §2.1.

Si dimostra il valore di $cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right)$ nel caso in cui lo stimatore calibrato con vincoli campionari utilizza una sola variabile ausiliaria da una fonte esterna, $\hat{\hat{X}}$, in base alle ipotesi sul coefficiente di regressione (stocastico o costante) e la correttezza della stima (corretta o distorta) della variabile ausiliaria da indagine campionaria dipendete.

$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = cov\left(\hat{Y}_{GREG}, \hat{B}\left(\hat{\hat{X}} - X\right)\right)$

Poiché:

$cov(X,Y) = E[XY] - E[X]E[Y]$

e, poiché le due quantità $\left(\hat{Y}_{GREG}, \hat{B}\left(\hat{\hat{X}} - X\right)\right)$

$E[XY] = E[X] + E[Y] - E[X]E[Y]$,

si ha che:

$cov(X,Y) = E[X] + E[Y] - 2E[X]E[Y]$

La $cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right)$ diventa dunque:

$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = E[\hat{Y}_{GREG}] + E\left[\hat{B}\left(\hat{\hat{X}} - X\right)\right] - 2\left(E[\hat{Y}_{GREG}]E\left[\left(\hat{B}\left(\hat{\hat{X}} - X\right)\right)\right]\right)$

$E[\hat{Y}_{GREG}]$ è il valore atteso dello stimatore di regressione generalizzato della variabile *Y* che, per popolazioni di grandi dimensioni, è corretto e coincide con il valore reale delle variabile nella popolazione: dunque è pari ad *Y*;

$E\left[\hat{B}\left(\hat{\hat{X}} - X\right)\right]$, svolgendo il prodotto e utilizzando la proprietà lineare del valore atteso, è uguale a:

$E\left[\hat{B}\hat{\hat{X}} - BX\right] = E\left[\hat{B}\hat{\hat{X}}\right] - E[\hat{B}X]$

X è il valore reale della variabile nella popolazione, è una costante e può essere messa fuori dal valore atteso;

$E\left[\hat{B}\hat{\hat{X}}\right]$, invece, è uguale, data la non completa indipendenza delle due quantità e come visto in precedenza, a:

$E\left[\hat{B}\hat{\hat{X}}\right] = E[\hat{B}] + E\left[\hat{\hat{X}}\right] - E[\hat{B}]E\left[\hat{\hat{X}}\right]$.

Quindi:

$E\left[\hat{B}\left(\hat{\hat{X}} - X\right)\right] = E[\hat{B}] + E\left[\hat{\hat{X}}\right] - E[\hat{B}]E\left[\hat{\hat{X}}\right] - X(E[\hat{B}])$.

Risultato 1. Se si considera $\hat{B}$ una variabile stocastica, sostituendo semplicemente le espressioni ricavate in precedenza, si ha che la covarianza di $\hat{Y}_{GREG}$ e $bias(\hat{Y}_{\widehat{PV}})$ è:

$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) =$

$= Y + E[\hat{B}] + E\left[\hat{\hat{X}}\right] - E[\hat{B}]E\left[\hat{\hat{X}}\right] - X(E[\hat{B}]) - 2\left(Y\left(E[\hat{B}] + E\left[\hat{\hat{X}}\right] - E[\hat{B}]E\left[\hat{\hat{X}}\right] - X(E[\hat{B}])\right)\right)$

Essendo $\hat{B}$ uno stimatore corretto di *B*, nel caso in cui $\hat{\hat{X}}$ è corretto $\left(E\left[\hat{\hat{X}}\right] = X\right)$ si ha che:

$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = Y + B + X - 2BX - 2YB - 2YX + 4YBX$

Nel caso in cui, invece, $\hat{\hat{X}}$ è non corretto $\left(E\left[\hat{\hat{X}}\right] = X + b\right)$ la covarianza di $\hat{Y}_{GREG}$ e $\hat{X}$ diventa:

$$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = Y + B + X + b - 2BX - Bb - 2YB - 2YX - 2Yb + 4YBX + 2YBb$$

Risultato 2. Poiché si può dimostrare empiricamente che B è stabile nel tempo, si può facilmente considerare una costante, dunque si avrà che:

$$E[\hat{B}X] = BX$$

e

$$E\left[\hat{B}\hat{\bar{X}}\right] = B\left(E\left[\hat{\bar{X}}\right]\right).$$

Dunque:

$$E\left[\hat{B}\left(\hat{\bar{X}} - X\right)\right] = B\left(E\left[\hat{\bar{X}}\right]\right) - BX.$$

Sostituendo questi valori nell'espressione precedente si ha che la covarianza di $\hat{Y}_{GREG}$ e $\hat{\bar{X}}$ è:

$$cov\left(\hat{Y}_{GREG}, \hat{B}\left(\hat{\bar{X}} - X\right)\right) = Y + B\left(E\left[\hat{\bar{X}}\right]\right) - BX - 2\left(Y\left(B\left(E\left[\hat{\bar{X}}\right]\right) - BX\right)\right)$$

Nel caso in cui $\hat{\bar{X}}$ è corretto $\left(E\left[\hat{\bar{X}}\right] = X\right)$ si ha che:

$$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = Y$$

Nel caso in cui, invece, $\hat{\bar{X}}$ è non corretto $\left(E\left[\hat{\bar{X}}\right] = X + b\right)$ la covarianza di A_1 e A_2 diventa:

$$cov\left(\hat{Y}_{GREG}, bias(\hat{Y}_{\widehat{PV}})\right) = Y + Bb - 2YBb$$

Bibliografia

Ballin, M., Falorsi, P. D. e Russo, A. (2000). Condizioni di Coerenza e Metodi di Stima per le Indagini Campionarie sulle Imprese. *Rivista di Statistica Ufficiale.* 2000, n. 2, pp. 31-52.

Bethlehem, J. G. e Keller, W. J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics.* 1987, vol. 3, n. 2, pp. 141-153.

Ceccarelli C. e A. Cutillo (2007). Il Trattamento della Mancata Risposta Totale nell'Indagine Eu-Silc: Una Valutazione Tramite una Misura del Cambiamento. *Congiuntura*. Udine: CREF, 1° trimetre, pp. 91-112.

Ceccarelli, C., Di Marco, M. e Rinaldelli, C. (2008). L'indagine Europea sui Redditi e le Condizioni di Vita delle Famiglie (It-silc). *Metodi e Norme.* Istat, 2008, n. 37.

Cicchitelli, G., Herzel, A. e Montanari G. E. (1992). *Il Campionamento Statistico.* Bologna: Il Mulino, 1992.

Deville, J.C e Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association.* Giugno, 1992, vol. 87, n. 418, pp. 376-382.

Deville, J. C., Särndal, C. E. e Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association.* Settembre, 1993, vol. 88, n. 243, pp. 1013-1020.

Deville, J.C. (1999). Simultaneous Calibration of Several Surveys.*Proceedings of statistic Canada Symposium 99. Combining Data from Different Sources.* Ottawa: Statistics Canada, Maggio 1999, pp. 207-212.

Estevao, V., Hidiroglou, M. A. e Särndal, C. E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics.* 1995, vol. 11, n. 2, pp. 181-204.

Gazzelloni, S. (2006). La Rilevazione sulle Forze di Lavoro: Contenuti, Metodologie, Organizzazione. *Metodi e norme.* Istat, 2006, n. 32.

Lemaître, G. e J., Dufour. (1987). An integrated method for Weighting Persons and Families. *Survey Methodology.* Dicembre, 1987, vol. 13, n. 2, pp. 199-207.

Pagliuca, D. (2005a). *Genesees V. 3.0, Funzione Riponderazione - Manuale Utente e Aspetti Metodologici.* Roma: Istat, 2005, Appendice (pp.141-214).

Pagliuca, D. (2005b). *Genesees V. 3.0, Funzione Stime ed Errori - Manuale Utente e Aspetti Metodologici.* Roma: Istat, 2005, Appendice (pp.189-246).

Singh, A.C e Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. *Survey Methodology.* Dicembre, 1996, vol. 22 , n. 2, pp. 107-115.

Wright, R. L. (1983). Finite Population Sampling with Multivariate Auxiliary Information . *Journal of the American Statistical Association.* Dicembre, 1983, vol. 78, n. 384, pp. 879-883.

Zannella, F. (1989). *Manuale di Tecniche di Indagine, 5 - Tecniche di Stima della Varianza Campionaria.* Roma: Istat, 1989.