quaderni di discussione

Sulla presentazione degli errori di campionamento mediante modelli. Il metodo dei modelli regressivi

ALDO RUSSO

istat

I quaderni di discussione sono a circolazione ristretta e non impegnano la responsabilità dell'ISTAT ma riflettono solo il punto di vista degli autori. Non possono, quindi, essere citati e fatti circolare senza il permesso degli autori.

Le richieste vanno indirizzate a :
«ISTAT - Centro Documentazione - Dr.ssa Borgnino-Valenzano
Via Balbo, 16 - 00100 - ROMA

N. 87. 04

Sulla presentazione degli errori di campionamento mediante modelli.
Il metodo dei modelli regressivi

ALDO RUSSO

MARZO 1987

Aldo RUSSO

Direttore di sezione, Reparto Studi,

Progetto I: Studio dei campioni

Indice

Riassunto

Scopo del lavoro é quello di presentare un approccio globale, basato sull'impiego dei modelli regressivi, che può costituire un valido punto di riferimento per chiunque voglia accostarsi alla problematica della presentazione degli errori di campionamento mediante modelli.

Pur cercando di sviluppare un discorso continuo, il lavoro si divide in sostanza in tre parti di cui la prima é dedicata ad alcune considerazioni teoriche tendenti ad evidenziare i fondamenti statistici su cui riposa l'approccio in esame; la seconda parte esamina alcuni metodi di stima per modelli lineari e non lineari nei parametri; le pagine conclusive, infine, intendono illustrare le varie tappe da percorrere per operare in modo statisticamente fondato nel caso di indagini reali.

1 - Introduzione

Durante lo svolgimento di ogni indagine campionaria esiste la possibilità di commettere errori. Una prima suddivisione, accettata da buona parte della letteratura statistica sullo argomento, conduce a distinguere gli errori di campionamento, dovuti alla natura parziale della rilevazione, da quelli di misura (o di risposta), derivanti da numerosi e spesso incontrollabili fattori di disturbo (Cfr. nota I).

La piena coscienza della loro esistenza e della loro importanza induce lo statistico, al fine di minimizzarne la portata, a prendere decisioni e provvedimenti che investono da una parte gli aspetti del disegno di campionamento (Cfr. nota II), e dall'altra le caratteristiche delle principali operazioni sia preparatorie sia esecutive concernenti l'indagine (Cfr. nota III); decisioni e provvedimenti che, tuttavia, non sono sempre sufficienti a ridurre apprezzabilmente l'entità dell'errore globale (Cfr. nota IV).

D'altra parte l'esperienza spesso mostra che, completata la rilevazione ed effettuate le tradizionali operazioni di revisione, rettifica e di analisi critica interna dei dati, i risultati dell'indagine vengono considerati pronti per essere passati agli utilizzatori, assumendo, più o meno esplicitamente, che le entità degli errori campionari e degli errori di misura possano considerarsi limitate e comunque tali da non compromettere seriamente la qualità dei risultati stessi.

Ma questa assunzione dovrebbe sempre essere onestamente verificata attraverso il calcolo dell'errore campionario e la valutazione almeno degli errori di misura più pericolosi, in modo da poter apprezzare l'entità dell'errore globale.

E' stato altresì auspicato, sia in sede internazionale (Cfr. [2], [41]) sia da parte di studiosi di statistica (Cfr. [7], [12]) - con lo scopo di fornire agli utilizzatori elementi obiettivi di giudizio indispensabili per la corretta interpretazione dei dati ottenuti - che le relazioni finali sui risultati delle indagini abbiano un capitolo speciale in cui siano descritte la metodologia di campionamento adottata, le modalità di raccolta delle informazioni e le formule utilizzate per il riporto all'universo dei dati, e che riporti gli errori campionari e di misura ed alcune statistiche ausiliarie utili nell'inferenza e nell'analisi statistica (Cfr. nota V).

Ed é in questo spirito che l'ISTAT, a partire dai primi anni '80, ha avviato un programma di studi per affrontare sistematicamente e scientificamente il calcolo dell'errore di campionamento; solo più recentemente, invece, é stata presa l'iniziativa di studiare e di valutare almeno la parte più grossa dell'errore di misura (Cfr. nota VI).

Limitatamente all'aspetto campionario, una informazione completa sul livello di precisione dei risultati richiederebbe la specificazione degli errori campionari di tutte le stime pubblicate. Ciò tuttavia - per la complessità dei disegni campionari generalmente adottati (Cfr. nota VII) - non é proponibile in quanto comporterebbe un elevato numero di elaborazioni meccanografiche, un appesantimento delle tavole di pubblicazione e di conseguen

za tempi di calcolo e costi più elevati.

Tali difficoltà hanno ben presto offerto lo spunto per introdurre alcuni metodi approssimati che agevolano notevolmente il calcolo degli errori campionari (Cfr.nota VIII) ed idonei modelli che consentono di esporre in forma concisa i suddetti errori(Cfr.nota IX).

In precedenti lavori(Cfr.[27],[30],[31],[32],[38]) abbiamo già utilizzato il metodo dello sviluppo in serie di Taylor di una funzione (detto anche metodo di linearizzazione) e il metodo dei modelli regressivi,rispettivamente,per la determinazione e la presentazione concisa degli errori campionari.

In questa nota,limitatamente al metodo dei modelli regressivi,approfondiamo i temi ivi abbozzati sviluppando anche talune considerazioni teoriche al fine di giustificare sul piano metodologico l'impiego di tale metodo.

2 - Il metodo dei modelli regressivi

2.1 - Premessa

Supponiamo di aver effettuato un'indagine ed indichiamo con:

$$G_s = (\hat{X}_1, \ldots\ldots\ldots, \hat{X}_i, \ldots\ldots\ldots, \hat{X}_s)$$

un gruppo di s stime campionarie.

Indichiamo inoltre, rispettivamente, con:

$$X_1, \ldots\ldots\ldots, X_i, \ldots\ldots\ldots, X_s$$

$$V(\hat{X}_1), \ldots\ldots, V(\hat{X}_i), \ldots\ldots, V(\hat{X}_s)$$

$$\in^2(\hat{X}_1), \ldots\ldots, \in^2(\hat{X}_i), \ldots\ldots, \in^2(\hat{X}_s)$$

i corrispondenti valori attesi, varianze e varianze relative, queste ultime definite da :

$$\in^2(\hat{X}_i) = \frac{V(\hat{X}_i)}{X_i^2} \qquad (i=1, \ldots, s) \qquad (1)$$

Il fine che persegue il metodo dei modelli regressivi é la determinazione di una opportuna equazione matematica mediante la quale é possibile pervenire ad una stima del livello di precisione delle s stime $\hat{X}_1, \ldots, \hat{X}_i, \ldots, \hat{X}_s$.

L'avvio alla trattazione di tale metodo può farsi derivare, come é stato suggerito da taluni autori, da un'ipotesi fondamentale: quella cioé che, nel gruppo G_s precedentemente definito, il comportamento della varianza relativa (o dell'errore relativo $\in(\hat{X}_i)$) dipenda soltanto dall'ampiezza del valore atteso; simbolicamente un siffatto comportamento può essere espresso da una relazione funzionale del tipo seguente:

$$\epsilon^2(\hat{X}) = f(X, a_1, a_2, \ldots, a_q, u) \qquad (2)$$

in cui $a_1, a_2, \ldots, a_q$ sono delle costanti ed u una perturbazione (o errore) stocastica.

In pratica la relazione (2) viene sostituita dall'analoga relazione operativa:

$$\hat{\epsilon}^2(\hat{X}) = f(\hat{X}, a_1, a_2, \ldots, a_q, u) \qquad (3)$$

in cui $\hat{\epsilon}^2(\hat{X})$ rappresenta una stima di $\epsilon^2(\hat{X})$ definita da:

$$\hat{\epsilon}^2(\hat{X}) = \frac{\hat{V}(\hat{X})}{\hat{X}^2} \qquad (4)$$

essendo $\hat{V}(\hat{X})$ una stima di $V(\hat{X})$; la situazione può essere rappresentata come nella figura I.

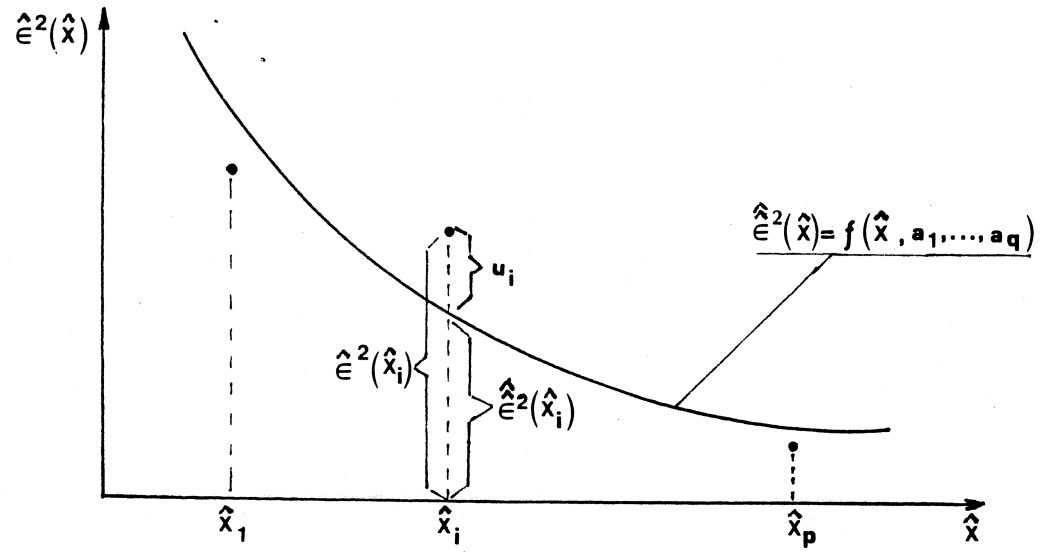

Fig.1

La stima dei parametri $a_1, a_2, \ldots, a_q$ viene ottenuta adattando il modello (3) alla nuvola di punti $N\left(\hat{X}_i, \hat{\epsilon}^2(\hat{X}_i)\right)$, formata dalle p stime:

$$G_p = \left(\hat{X}_1, \ldots\ldots\ldots, \hat{X}_i, \ldots\ldots\ldots, \hat{X}_p\right)$$

costituenti un sottoinsieme di G_s e dalle corrispondenti varian ze relative:

$$\hat{\in}^2(\hat{X}_1), \ldots\ldots, \hat{\in}^2(\hat{X}_i), \ldots\ldots, \hat{\in}^2(\hat{X}_p)$$

Disponendo del modello stimato:

$$\hat{\in}^2(\hat{X}) = f(\hat{X}, \hat{a}_1, \hat{a}_2, \ldots\ldots, \hat{a}_q, e) \qquad (5)$$

in cui $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q$ indicano rispettivamente le stime dei pa rametri incogniti $a_1, a_2, \ldots, a_q$ ed e rappresenta il residuo de finito dalla differenza:

$$e = \hat{\in}^2(\hat{X}) - \hat{\hat{\in}}^2(\hat{X})$$

dove $\hat{\hat{\in}}^2(\hat{X})$ indica il valore dell'ordinata corrispondente al va lore $\hat{X}$,é possibile determinare per ciascuna stima $\hat{X}_i$ (i=I,.,s) una stima della corrispondente varianza relativa mediante la re lazione:

$$\hat{\hat{\in}}^2(\hat{X}_i) = f(\hat{X}_i, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) \qquad (6)$$

Graficamente questa situazione é illustrata nella fi gura 2.

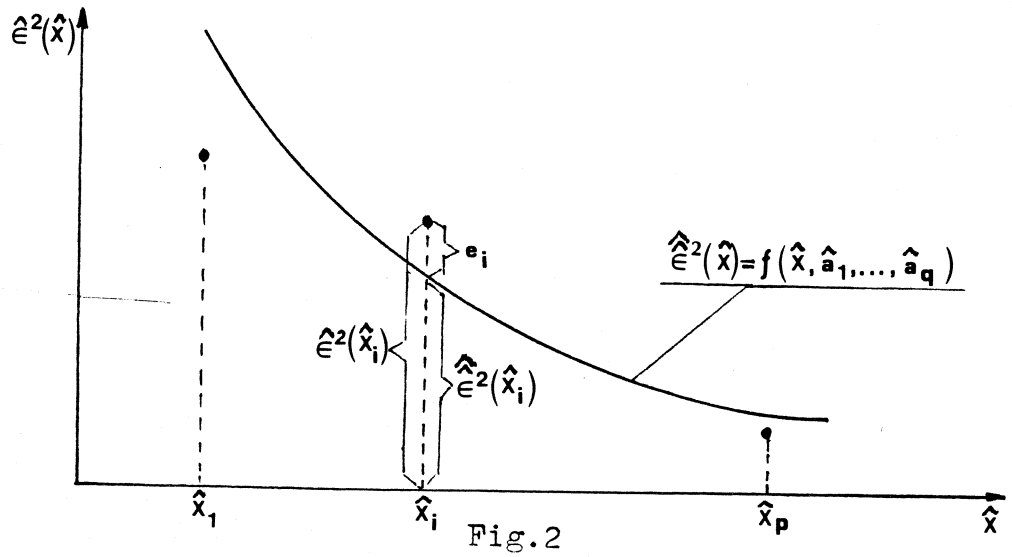

Fig.2

Dalla (6) possono poi ricavarsi l'errore relativo ed as soluto rispettivamente forniti dalle espressioni:

$$\hat{\hat{\epsilon}}(\hat{X}_i) = \sqrt{f(\hat{X}_i, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)} \qquad (7)$$

$$\hat{\hat{\sigma}}(\hat{X}_i) = \hat{\hat{\epsilon}}(\hat{X}_i) \cdot \hat{X}_i \qquad (8)$$

Nei rapporti riportanti i risultati di un'indagine viene tuttavia adottato, per ragioni di semplicità e praticità, un approccio basato sull'utilizzazione di un prospetto del tipo seguente:

$\hat{X}'_i$	$\hat{\hat{\epsilon}}(\hat{X}'_i)$
$\hat{X}'_1$	$\hat{\hat{\epsilon}}(\hat{X}'_1)$
.	.
.	.
$\hat{X}'_j$	$\hat{\hat{\epsilon}}(\hat{X}'_j)$
.	.
.	.
$\hat{X}'_J$	$\hat{\hat{\epsilon}}(\hat{X}'_J)$

in cui nella prima e nella seconda colonna sono indicati, rispettivamente, alcuni particolari livelli di stima (Cfr.nota X) e i corrispondenti valori dell'errore relativo ricavati attraverso la (7), ponendo $\hat{X}'_j$ al posto di $\hat{X}_i$.

Il calcolo dell'errore relativo corrispondente alla generica stima $\hat{X}_i \in G_s$ può essere effettuato secondo i due procedimenti seguenti:

i) cercando nella prima colonna del suddetto prospetto il livello di stima che più si avvicina al valore $\hat{X}_i$; l'errore relati

vo $\hat{\hat{\in}}(\hat{X}_i)$ si troverà sulla stessa riga della seconda colonna;

ii) indicando con $\hat{X}'_{i-1}$ ed $\hat{X}'_{i+1}$ i valori delle stime entro i quali é compreso $\hat{X}_i$ e con $\hat{\hat{\in}}(\hat{X}'_{i-1})$ e $\hat{\hat{\in}}(\hat{X}'_{i+1})$ i corrispondenti errori relativi, riportati nel prospetto, il valore $\hat{\hat{\in}}(\hat{X}_i)$ si ricava per interpolazione lineare tramite la relazione:

$$\hat{\hat{\in}}(\hat{X}_i) = \hat{\hat{\in}}(\hat{X}'_{i-1}) - \frac{\hat{\hat{\in}}(\hat{X}'_{i-1}) - \hat{\hat{\in}}(\hat{X}'_{i+1})}{\hat{X}'_{i+1} - \hat{X}'_{i-1}} (\hat{X}_i - \hat{X}'_{i-1})$$

Ci sembra utile sottolineare che l'impiego del metodo dei modelli regressivi, richiedendo il calcolo delle varianze relative limitatamente ad un sottoinsieme di stime di dimensione generalmente molto minore di quella dell'insieme G_s, consente una considerevole riduzione dei tempi di calcolo e dei costi rispetto al criterio di specificare accanto ad ogni stima pubblicata il corrispondente errore di campionamento.

Nel corso di quest'ultimo quindicennio i modelli che si sono dimostrati più idonei a descrivere il legame espresso dalla (2) sono rappresentati dalle seguenti forme (Cfr.nota XI):

a) $$\in^2(\hat{X}) = a_1 + \frac{a_2}{X} + u$$

b) $$\log \in (\hat{X}) = a_1 + a_2 \log X + u$$

c) $$\in^2 (\hat{R}) = \frac{a_2}{Y} \frac{1 - R}{R} + u$$

d) $$\in^2(\hat{R}) = a_1 + \frac{a_2}{R\,Y} + \frac{a_3}{Y} + u$$

Nelle applicazioni pratiche i suddetti modelli vengono ad essere sostituiti dagli analoghi modelli operativi definiti rispettivamente dalle forme:

a') $$\hat{\epsilon}^2(\hat{X}) = a_1 + \frac{a_2}{\hat{X}} + u$$

b') $$\log \hat{\epsilon}(\hat{X}) = a_1 + a_2 \log \hat{X} + u$$

c') $$\hat{\epsilon}^2(\hat{R}) = \frac{a_2}{\hat{Y}} \frac{I - \hat{R}}{\hat{R}} + u$$

d') $$\hat{\epsilon}^2(\hat{R}) = a_1 + \frac{a_2}{\hat{R}\ \hat{Y}} + \frac{a_3}{\hat{Y}} + u$$

avendo indicato con $\hat{X}$ ed $\hat{R}$ le stime di X ed R e con $\hat{\epsilon}^2(\hat{X})$ ed $\hat{\epsilon}^2(\hat{R})$ le stime di $\epsilon^2(\hat{X})$ ed $\epsilon^2(\hat{R})$.

I modelli a') e b') riguardano il caso delle stime di frequenze assolute (Cfr. [32], [46]); i rimanenti modelli si riferiscono invece al caso di stime rapporto (Cfr. [46], [49]).

Nei successivi paragrafi 2.2, 2.3 e 2.4 illustreremo i suddetti modelli mettendone in luce la validità ai fini di una loro utilizzazione per descrivere e rappresentare sinteticamente la relazione varianza relativa-stima.

2.2 - Stime di frequenze assolute

Immaginiamo di aver effettuato un'indagine basata su disegno campionario complesso.

Per maggiore comprensione e chiarezza della trattazione supponiamo, ad esempio, che si sia fatto ricorso ad un campione a due stadi, stratificato al primo stadio, con selezione delle unità in ciascuno stadio senza reimmissione e con probabilità uguale.

Supponiamo inoltre che le unità primarie siano costituite dai comuni e quelle secondarie dalle famiglie, all' interno delle quali tutti i componenti appartenenti alla popolazione oggetto d'indagine vengono intervistati.

Indichiamo poi con:

h	indice di strato $(h=I,2,.....,H)$
i	indice di unità primaria (comune)
j	indice di unità secondaria (famiglia)
N_h	numero di comuni- universo nello strato h
n_h	numero di comuni- campione nello strato h
M_{hi}	numero di famiglie-universo nel comune i dello strato h
m_{hi}	numero di famiglie-campione nel comune i dello strato h
P_{hij}	numero di componenti della famiglia j del comune i dello strato h
P	ampiezza complessiva della popolazione oggetto d'indagine
p	numero complessivo di persone intervistate
X_{hij}	numero di persone aventi un dato carattere nella famiglia j del comune i dello strato h

X_{hi}	numero di persone aventi un dato carattere nel comune i dello strato h
X_h	numero di persone aventi un dato carattere nello strato h
X	numero di persone aventi un dato carattere nella popolazione
$\Pi = X / P$	frazione di persone aventi un dato carattere nella popolazione

Ciò premesso, é possibile dimostrare (Cfr. [48]) che nel campionamento a due stadi, con stratificazione delle unità primarie e con selezione delle unità secondo il meccanismo probabilistico già descritto, una stima corretta del parametro X é fornita dall'espressione:

$$\hat{X} = \sum_{1}^{H}{}_{h} \sum_{1}^{n_h}{}_{i} \sum_{1}^{m_{hi}}{}_{j} \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} X_{hij} = \tag{9}$$

$$= \sum_{1}^{H}{}_{h} \sum_{1}^{n_h}{}_{i} \sum_{1}^{m_{hi}}{}_{j} K_{hi} X_{hij}$$

in cui:

$$K_{hi} = \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \tag{10}$$

rappresenta il peso attribuito alle m_{hi} famiglie-campione del comune i dello strato h.

La varianza campionaria della stima $\hat{X}$ può essere espressa come prodotto della varianza di un campione casuale semplice di numerosità p per il fattore deff (effetto del disegno di campionamento, Cfr. nota XII):

$$V(\hat{X}) = P^2 \frac{P - p}{P - 1} \frac{\Pi(1-\Pi)}{p} \text{ deff} \qquad (11)$$

Dalla (11) segue che la varianza relativa di $\hat{X}$ é data da:

$$\epsilon^2(\hat{X}) = \frac{V(\hat{X})}{X^2} = \frac{P^2}{X^2} \frac{P - p}{P - 1} \frac{\Pi(1-\Pi)}{p} \text{ deff} \qquad (12)$$

che, attraverso alcuni semplici passaggi che omettiamo di trascrivere, può porsi nella forma:

$$\epsilon^2(\hat{X}) = - \frac{K}{P} \text{ deff} + K \text{ deff } \frac{1}{X} \qquad (13)$$

in cui si é posto:

$$K = \frac{P}{p} \frac{P - p}{P - 1} \qquad (14)$$

Se deff é costante (o approssimativamente tale) nel gruppo costituito dalle s stime $\hat{X}_1, \ldots, \hat{X}_i, \ldots, \hat{X}_s$, il modello:

$$\epsilon^2(\hat{X}) = a_1 + \frac{a_2}{X} + u \qquad (15)$$

può essere utilizzato per stimare la varianza nel gruppo stesso.

Un modello alternativo di (15) può derivarsi con altro ragionamento basato sull'errore campionario relativo.

Dalla (13) si ricava che:

$$\epsilon(\hat{X}) = \sqrt{\frac{K \text{ deff}}{X} \left(1 - \frac{X}{P} \right)} \qquad (16)$$

Applicando i logaritmi ad ambo i membri della (16) si ottiene:

$$\log \in(\hat{X}) = \frac{1}{2} \log K \quad deff \quad - \frac{1}{2} \log X + \frac{1}{2}\left[1 - \frac{X}{P}\right] \qquad (17)$$

Per la presenza del terzo termine a secondo membro, la (17)-anche assumendo deff = costante - non é lineare in $\log \in(\hat{X})$ e log X; per piccoli valori di X, tuttavia, il contributo dovuto a tale termine diviene trascurabile.

In tale circostanza e sotto l'ipotesi che deff sia costante o approssimativamente tale nel gruppo costituito dalle stime $\hat{X}_1, \ldots, \hat{X}_i, \ldots, \hat{X}_s$, il modello:

$$\log \in(\hat{X}) = a_1 + a_2 \log X + u \qquad (18)$$

consente di stimare la varianza nel gruppo stesso (Cfr.nota XIII).

2.3 - Stime rapporto in cui il numeratore é un sottoinsieme del del denominatore

Consideriamo un'indagine eseguita su n unità costituenti un campione estratto senza reimmissione e con probabilità uguale da una popolazione di N unità.

Indichiamo poi con:

i	indice di unità
X_i	variabile che assume il valore 1 se l'unità i presenta entrambi gli attributi x ed y e 0 altrimenti
Y_i	variabile che assume il valore 1 se l'unità i presenta l'attributo y e 0 altrimenti

$X = \sum_1^N{}_i X$ numero di unità della popolazione che presentano entrambi gli attributi x ed y

$Y = \sum_1^N{}_i Y$ numero di unità della popolazione che presentano l'attributo y

$R = X/Y$ proporzione delle unità che presentano gli attributi x ed y fra quelle che possiedono solo l'attributo y

La stima di R é data da:

$$\hat{R} = \frac{\hat{X}}{\hat{Y}} \tag{19}$$

in cui:

$$\hat{X} = \frac{N}{n} \sum_1^N{}_i X_i \, \delta_i \tag{20}$$

$$\hat{Y} = \frac{N}{n} \sum_1^N{}_i Y_i \, \delta_i \tag{21}$$

rappresentano, rispettivamente, le stime corrette di X e Y, essendo δ_i una variabile che assume il valore 1 se la generica unità é inclusa nel campione e 0 altrimenti.

La varianza relativa di $\hat{R}$ é fornita dall'espressione:

$$\epsilon^2(\hat{R}) = \frac{V(\hat{R})}{R^2} = \frac{V(\hat{X})}{X^2} + \frac{V(\hat{Y})}{Y^2} - 2\,\frac{C(\hat{X}\,,\,\hat{Y})}{X\,Y} \tag{22}$$

Ciò premesso dimostriamo che nelle condizioni sopra poste si ha la relazione:

$$\frac{C(\hat{X}\,,\hat{Y})}{X\,Y} = \frac{V(\hat{Y})}{Y^2} \tag{23}$$

A tale scopo, determiniamo in primo luogo un'espressione esplicita del primo membro della (23).

Dalla nota relazione:

$$C(\hat{X}, \hat{Y}) = E(\hat{X}, \hat{Y}) - E(\hat{X})\, E(\hat{Y}) = E(\hat{X}, \hat{Y}) - X\, Y \qquad (24)$$

si ottiene:

$$E(\hat{X}, \hat{Y}) = C(\hat{X}, \hat{Y}) + X\, Y = N^2 \frac{N-n}{N-1} \frac{C(X, Y)}{n} + X\, Y =$$

$$= \frac{N^2}{n} \frac{N-n}{N-1} \left[\frac{1}{N} \sum_{1}^{N}{}_{i} X_i\, Y_i - \frac{X}{N} \frac{Y}{N} \right] + X\, Y =$$

$$= \frac{N}{n} \frac{N-n}{N-1} \sum_{1}^{N}{}_{i} X_i\, Y_i - \frac{N-n}{N-1} \frac{X\, Y}{n} + X\, Y = \qquad (25)$$

$$= \frac{N}{n} \frac{N-n}{N-1} \sum_{1}^{N}{}_{i} X_i\, Y_i + \frac{N(n-1)}{n(N-1)} X\, Y =$$

$$= \frac{N}{n} \frac{N-n}{N-1} X + \frac{N(n-1)}{n(N-1)} X\, Y =$$

$$= \frac{N}{n} X \left[\frac{N-n}{N-1} + \frac{n-1}{N-1} Y \right]$$

In base al risultato espresso dalla (25) la (24) può porsi nella forma:

$$C(\hat{X}, \hat{Y}) = \frac{N}{n} X \left[\frac{N-n}{N-1} + \frac{n-1}{N-1} Y \right] - X\, Y \qquad (26)$$

Pertanto il primo membro della (23) può scriversi:

$$\frac{C(\hat{X},\hat{Y})}{X\,Y} = \frac{N}{n}\,\frac{1}{Y}\left[\frac{N-n}{N-1} + \frac{n-1}{N-1}\,Y\right] - 1 \qquad (27)$$

Poniamo ora in forma diversa il secondo membro della (23). Dalla nota relazione:

$$V(\hat{Y}) = E(\hat{Y}^2) - \left[E(\hat{Y})\right]^2 = E(\hat{Y}^2) - Y^2 \qquad (28)$$

si ricava che:

$$E(\hat{Y}^2) = V(\hat{Y}) + Y^2 =$$

$$= N^2\frac{N-n}{N-1}\,\frac{V(Y)}{n} + Y^2 =$$

$$= \frac{N^2}{n}\,\frac{N-n}{N-1}\left[\frac{1}{N}\sum_1^N{}_i\, Y_i^2 - \frac{Y^2}{N^2}\right] + Y^2 = \qquad (29)$$

$$= \frac{N}{n}\,\frac{N-n}{N-1}\sum_1^N{}_i\, Y_i^2 - \frac{N-n}{N-1}\,\frac{Y^2}{n} + Y^2 =$$

$$= \frac{N}{n}\,\frac{N-n}{N-1}\sum_1^N{}_i\, Y_i^2 + \frac{N}{n}\,\frac{n-1}{N-1}\,Y^2 =$$

$$= \frac{N}{n}\,\frac{N-n}{N-1}\,Y + \frac{N}{n}\,\frac{n-1}{N-1}\,Y^2 =$$

$$= \frac{N}{n}\,Y\left[\frac{N-n}{N-1} + \frac{n-1}{N-1}\,Y\right]$$

Se si tiene conto della (29) nella (28) si ottiene:

$$V(\hat{Y}) = \frac{N}{n} Y \left[\frac{N-n}{N-1} + \frac{n-1}{N-1} Y \right] - Y^2 \qquad (30)$$

da cui discende che il primo membro della (23) assume la forma

$$\frac{V(\hat{Y})}{Y^2} = \frac{N}{n} \frac{1}{Y} \left[\frac{N-n}{N-1} + \frac{n-1}{N-1} Y \right] - 1 \qquad (31)$$

Essendo la (31) formalmente coincidente con l'espressione (27) resta così dimostrata la relazione (23), in base alla quale la (22) diviene:

$$\epsilon^2(\hat{R}) = \frac{V(\hat{X})}{X^2} + \frac{V(\hat{Y})}{Y^2} - 2 \frac{V(\hat{Y})}{Y^2} = \qquad (32)$$

$$= \frac{V(\hat{X})}{X^2} - \frac{V(\hat{Y})}{Y^2} = \epsilon^2(\hat{X}) - \epsilon^2(\hat{Y})$$

Se le varianze relative a secondo membro della (32) sono esprimibili mediante la relazione (I5), ossia se:

$$\epsilon^2(\hat{X}) = a_1 + \frac{a_2}{X} + u' \qquad \epsilon^2(\hat{Y}) = a_1 + \frac{a_2}{Y} + u'' \qquad (33)$$

si ricava immediatamente che la (32) può porsi nella forma:

$$\epsilon^2(\hat{R}) = a_2 \left(\frac{1}{X} - \frac{1}{Y} \right) + u = \frac{a_2}{Y} \frac{1-R}{R} + u \qquad (34)$$

avendo posto : $u' - u'' = u$

che, sotto la condizione molto restrittiva che $\hat{X}$ ed $\hat{Y}$ siano desunte da un campione casuale semplice, rappresenta il solo model

lo correttamente applicabile per stimare la varianza relativa di $\hat{R}$ (Cfr.nota XIV).

Nella pratica campionaria tuttavia il modello (34) viene efficacemente utilizzato anche nel caso di indagini basate su disegni campionari più complessi di quello casuale semplice, per i quali, a tutt'oggi, non é stata ancora dimostrata la validità della relazione (23), mediante la quale abbiamo derivato il modello (34); peraltro, l'abbondante documentazione sperimentale di cui si dispone offre elementi di giudizio che confortano l'uso del modello (34) anche per indagini condotte con campioni complessi.

2.4 - Stime rapporto in cui il numeratore non é un sottoinsieme del denominatore

Supponiamo di aver effettuato un'indagine campionaria basata su un disegno complesso che, senza alcuna perdita di generalità, possiamo supporre del tipo descritto al precedente paragrafo 2.2 .

Indichiamo poi con:

X — numero di persone aventi l'attributo x nella popolazione

Y — numero di persone aventi l'attributo y nella popolazione

$R = X/Y$ — rapporto tra i parametri X ed Y basato sull'assunzione che X non sia un sottoinsieme di Y

Siano inoltre:

$$\hat{X} = \sum_{1}^{H}{}_{h} \sum_{1}^{n_h}{}_{i} \sum_{1}^{m_{hi}}{}_{j} K_{hi} \; X_{hij} \qquad (35)$$

$$\hat{Y} = \sum_{1}^{H}{}_{h} \sum_{1}^{n_h}{}_{i} \sum_{1}^{m_{hi}}{}_{j} K_{hi} \quad Y_{hij} \qquad (36)$$

le stime corrette, rispettivamente, di X ed Y avendo indicato con:

X_{hij} il numero di persone aventi l'attributo x nella famiglia j del comune i dello strato h

Y_{hij} il numero di persone aventi l'attributo y nella famiglia j del comune i dello strato h

La stima di R é dunque fornita da:

$$\hat{R} = \frac{\hat{X}}{\hat{Y}} \qquad (37)$$

la cui varianza relativa é definita dall'espressione:

$$\epsilon^2(\hat{R}) = \frac{V(\hat{X})}{X^2} + \frac{V(\hat{Y})}{Y^2} - 2\,\frac{C(\hat{X}\,,\hat{Y})}{X\,Y} \qquad (38)$$

Nell'ipotesi che si possa supporre nulla o trascurabile la covarianza fra $\hat{X}$ ed $\hat{Y}$, la (38) assume la forma più semplice:

$$\epsilon^2(\hat{R}) = \frac{V(\hat{X})}{X^2} + \frac{V(\hat{Y})}{Y^2} \qquad (39)$$

Se, inoltre, le varianze relative di $\hat{X}$ ed $\hat{Y}$ sono esprimibili mediante la relazione (15), la (39) può scriversi:

$$\epsilon^2(\hat{R}) = a_1' + \frac{a_2'}{X} + a_1'' + \frac{a_2''}{Y} + u' + u'' \qquad (40)$$

Ponendo poi:

$$a_1 = a_1' + a_1'' \;,\; a_2 = a_2' \;,\; a_3 = a_2'' \;,\; u = u' + u'' \qquad (41)$$

la (40) assume la forma seguente:

$$\epsilon^2(\hat{R}) = a_1 + \frac{a_2}{X} + \frac{a_3}{Y} + u \qquad (42)$$

che può porsi nella forma equivalente (Cfr.nota XV):

$$\epsilon^2(\hat{R}) = a_1 + \frac{a_2}{R\,Y} + \frac{a_3}{Y} + u \qquad (43)$$

che rappresenta il modello comunemente adottato per stimare la varianza relativa di stime rapporto in cui il numeratore non é un sottoinsieme del denominatore (Cfr.nota XVI).

3. Procedimenti di stima dei parametri

3.1 - Metodo dei minimi quadrati

Il problema di stima dei parametri incogniti che caratterizzano il modello di regressione (3) precedentemente definito si risolve, in genere, con un procedimento che storicamente ha avuto una importanza notevolissima: il metodo dei minimi quadrati (Least Squares Methods, che abbrevieremo con LS).

Conformemente a tale metodo lo stimatore del vettore $v = (a_1, a_2, \ldots\ldots, a_q)$ dei parametri incogniti é quel vettore $\hat{v} = (\hat{a}_1, \hat{a}_2, \ldots\ldots, \hat{a}_q)$ che minimizza la somma dei quadrati degli scarti fra i "valori veri" della variabile dipendente, espressi dalle p varianze relative:

$$\hat{\in}^2(\hat{X}_i) \qquad (i=1,2,\ldots,p)$$

ed i p "valori calcolati" della variabile dipendente, espressi da:

$$\hat{\hat{\in}}^2(\hat{X}_i) \qquad (i=1,2,\ldots,p)$$

Lo stimatore $\hat{v}$ del vettore dei parametri incogniti v, ottenuto col metodo dei minimi quadrati, é pertanto la soluzione del seguente problema di minimizzazione:

$$\min\left(S(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) \right) = \min\left[\sum_1^p{}_i \left(\hat{\in}^2(\hat{X}_i) - \hat{\hat{\in}}^2(\hat{X}_i) \right)^2 \right] \qquad (44)$$

Come é noto, condizione necessaria (Cfr. nota XVII) per l'esistenza di un minimo é che si annullino le derivate parziali di $S(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)$ rispetto ad $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q$, ovvero:

$$\begin{cases} \dfrac{\partial(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)}{\partial \hat{a}_1} = 0 \\ \dfrac{\partial(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)}{\partial \hat{a}_2} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)}{\partial \hat{a}_q} = 0 \end{cases} \qquad (45)$$

Risolvendo il sistema di equazioni (45) - abitualmente indicato col termine di " equazioni normali " - si ottiene lo stimatore $\hat{v} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q)$.

Del metodo dei minimi quadrati esistono due casi particolari, ma di grande importanza. Il primo si ottiene quando si suppone che gli errori u_i soddisfano le condizioni seguenti:

$$\begin{cases} E(u_i) = 0 & \text{per } i=1,2,\ldots,p \\ E(u_i^2) = \sigma^2 & \text{per } i=1,2,\ldots,p \\ E(u_i, u_j) = 0 & \text{per } i \neq j \\ \hat{X} \text{ é una variabile deterministica} \end{cases} \qquad (46)$$

Sotto le suddette ipotesi, il modello é noto come " modello di regressione".

Aggiungiamo, ora, un breve commento sulle precedenti ipotesi per evidenziare i motivi per i quali vengono imposte e le conseguenze che implicano.

La prima ipotesi deriva dalla necessità che scarti positivi e negativi si compensino in media, coerentemente col carattere " accidentale " che si attribuisce alle u_i. In altri termini, le variabili casuali u_i non hanno nessuna influenza in media su $\hat{\epsilon}^2(\hat{X})$.

La seconda ipotesi indica che la varianza delle u_i rimane costante al variare delle osservazioni. Si suppone cioé che la variabilità non cresca né diminuisca al crescere o diminuire delle $\hat{\epsilon}^2(\hat{X}_i)$, (ipotesi di omoschedasticità).

La terza ipotesi implica che le variabili casuali u_i siano fra loro incorrelate, cioé abbiano covarianza nulla.

La quarta ipotesi considera i dati relativi alla variabile indipendente $\hat{X}$ come noti senza errori, quindi non soggetti a deviazioni di natura accidentale.

Il secondo caso particolare si ottiene quando, oltre alle (46), la $f(\hat{X}_i, a_1, a_2, \ldots, a_q)$ può essere espressa sotto forma di combinazione lineare dei parametri; in simboli:

$$f(\hat{X}, a_1, a_2, \ldots, a_q) = a_1 h_1(\hat{X}) + a_2 h_2(\hat{X}) + \cdots + a_q h_q(\hat{X}) \qquad (47)$$

in cui $h_1(\hat{X}), h_2(\hat{X}), \ldots, h_q(\hat{X})$, sono funzioni matematiche note della sola $\hat{X}$ non contenenti ulteriori parametri.

In questo caso si parla di modello di regressione lineare: il sistema (45) si riduce ad un sistema di equazioni lineari e gli stimatori $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q$, possono essere ottenuti esplicitamente (Cfr. nota XVIII).

L'importanza del modello di regressione lineare risiede nel fatto che gli stimatori LS sono (Cfr. nota XIX):

a) - lineari;

b) - non distorti e consistenti;

c) - hanno varianza minima nella classe degli stimatori lineari e non distorti.

Per tali motivi questi stimatori sono noti come BLUE (Best Linear Unbiased Estimators).

Occorre, naturalmente, porsi in atteggiamento critico rispetto a tutte le ipotesi del modello lineare di regressione, in quanto i dati spesso violano più di una di esse. Sarebbe cioé opportuno cercare, da un lato, di verificare se tali ipotesi sono plausibili, dall'altro, stabilire quali sono le conseguenze sui risultati ottenuti quando ciascuna di tali ipotesi non é valida ed elaborare delle procedure che ci permettano di migliorarne i risultati.

In questa nota affronteremo tale problematica -prendendo in esame le modifiche che devono essere introdotte- limitatamente ai seguenti due casi:

- non é verificata l'ipotesi di omoschedasticità;

- la funzione non é lineare nei parametri.

3.2 - Metodo dei minimi quadrati ponderati

Le numerose ricerche condotte sul problema dell'interpolazione degli errori campionari non sempre avallano l'ipotesi di omoschedasticità. In tali circostanze il metodo LS é tecnicamente applicabile, però non consente di pervenire ad una valutazione soddisfacente - come vedremo in seguito - dei parametri del modello. Sorge pertanto il problema di rimuoverla e di elaborare metodi che consentano una determinazione statisticamente più valida dei parametri del modello.

A tal fine consideriamo il modello lineare nei parametri:

$$\hat{\epsilon}^2(\hat{X}_i) = a_1 h_1(\hat{X}_i) + a_2 h_2(\hat{X}_i) + \cdots\cdots + a_q h_q(\hat{X}_i) + u_i \qquad (48)$$

e supponiamo che i p errori abbiano varianze $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$ diverse (ipotesi di eteroschedasticità), ferme restando tutte le altre ipotesi.

Sotto tali condizioni, é possibile dimostrare (Cfr. [20], [40]) che se nel modello di regressione lineare vi é eteroschedasticità e si stimano i parametri come se vi fosse omoschedasticità, le stime risultano non distorte e consistenti ma poco efficienti.

Si pone perciò il problema di superare tale inconveniente cercando di trasformare le variabili in modo da ottenere un nuovo modello con residui omoschedastici. Un tale procedimento é noto sotto il nome di " minimi quadrati ponderati " (Weighted Least Squares, che abbrevieremo con WLS).

Quando le varianze σ_i^2 sono note il modello (48) può essere trasformato nel seguente:

$$\frac{\hat{\epsilon}^2(\hat{X}_i)}{\sigma_i} = a_1\frac{h_1(\hat{X}_i)}{\sigma_i} + a_2\frac{h_2(\hat{X}_i)}{\sigma_i} + \ldots\ldots + a_q\frac{h_q(\hat{X}_i)}{\sigma_i} + \frac{u_i}{\sigma_i} \qquad (49)$$

Se poniamo:

$$\begin{cases} \hat{\epsilon}'^2(\hat{X}_i) = \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{\sigma_i} \\ h'_v(\hat{X}_i) = \dfrac{h_v(\hat{X}_i)}{\sigma_i} \quad \text{per } v=1,2,\ldots\ldots,q \\ u'_i = \dfrac{u_i}{\sigma_i} \end{cases} \qquad (50)$$

il modello (49) può riscriversi nella forma:

$$\hat{\epsilon}'^2(\hat{X}_i) = a_1h'_1(\hat{X}_i) + a_2h'_2(\hat{X}_i) + \ldots\ldots\ldots + a_qh'_q(\hat{X}_i) + u'_i \qquad (51)$$

che rappresenta un modello omoschedastico in quanto risulta:

$$E(u'_i)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2 = \frac{E(u_i^2)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1 \quad (i=1,\ldots,p) \qquad (52)$$

Quindi possiamo stimare i coefficienti con il metodo dei minimi quadrati e ovviamente le stime sono BLUE.

Quando $\sigma_1^2,\sigma_2^2,..,\sigma_p^2$ sono incognite e non é possibile stimarle occorre utilizzare un diverso procedimento.

A tal proposito osserviamo che, in alcuni casi, la diversa variabilità di u_i può essere attribuita ad una funzione-strettamente positiva-della variabile esplicativa $\hat{X}_i$; più precisamente si suppone che sia:

$$\sigma_i^2 = K^2 f^2(\hat{X}_i) \qquad (i=1,2,\ldots,p) \qquad (53)$$

per cui al posto della (49) si ha:

$$\frac{\hat{\epsilon}^2(\hat{X}_i)}{f(\hat{X}_i)} = a_1\frac{h_1(\hat{X}_i)}{f(\hat{X}_i)} + a_2\frac{h_2(\hat{X}_i)}{f(\hat{X}_i)} + \ldots\ldots\ldots + a_q\frac{h_q(\hat{X}_i)}{f(\hat{X}_i)} + \frac{u_i}{f(\hat{X}_i)} \qquad (54)$$

Il problema non presenta particolari difficoltà di soluzione. Infatti, le stime $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q$ di $a_1, a_2, \ldots, a_q$, si desumono imponendo il vincolo:

$$S'(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i) - \hat{\hat{\epsilon}}^2(\hat{X}_i)}{f(\hat{X}_i)} \right)^2 = \text{minimo} \qquad (55)$$

che può porsi nella forma (Cfr.nota XX):

$$S'(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) = \sum_1^p{}_i \left(\hat{\epsilon}^2(\hat{X}_i) - \hat{\hat{\epsilon}}^2(\hat{X}_i) \right)^2 p_i = \text{minimo} \qquad (56)$$

in cui:

$$p_i = \frac{1}{f^2(\hat{X}_i)} \qquad (57)$$

Vediamo ora come il procedimento appena descritto può essere utilizzato per risolvere il problema di stimare la relazione varianza relativa-stima, in presenza di eteroschedasticità.

Abbiamo già accennato, all'inizio di questo paragrafo, che gli studi sulla relazione in oggetto sembrano concludere che l'ipotesi di omoschedasticità non é realistica; molto verosimilmente, invece, la varianza residua cresce al diminuire della stima (Cfr.nota XXI). Poiché la varianza relativa cresce al diminuire della stima, é plausibile ipotizzare che sia:

$$\sigma_i^2 = K^2 \cdot \hat{\epsilon}^2(\hat{X}_i) \qquad (59)$$

di modo che il modello (54) diventa:

$$\frac{\hat{\epsilon}^2(\hat{X}_i)}{\hat{\epsilon}(\hat{X}_i)} = a_1\frac{h_1(\hat{X}_i)}{\hat{\epsilon}(\hat{X}_i)} + a_2\frac{h_2(\hat{X}_i)}{\hat{\epsilon}(\hat{X}_i)} + \ldots\ldots + a_q\frac{h_q(\hat{X}_i)}{\hat{\epsilon}(\hat{X}_i)} + \frac{u_i}{\hat{\epsilon}(\hat{X}_i)} \qquad (59)$$

e il vincolo (55) si scrive:

$$S'(\hat{a}_1,\hat{a}_2,\ldots,\hat{a}_q) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i) - \hat{\hat{\epsilon}}^2(\hat{X}_i)}{\hat{\epsilon}^2(\hat{X}_i)} \right)^2 = \text{minimo} \qquad (60)$$

3.3 - Metodo iterativo dei minimi quadrati ponderati

Accanto al metodo testé illustrato alcuni autori preferiscono considerare un procedimento iterativo dei minimi quadrati ponderati, basato sulla minimizzazione della seguente espressione:

$${}_sS''(\hat{a}_1,\hat{a}_2,\ldots,\hat{a}_q) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i) - {}_s\hat{\hat{\epsilon}}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right)^2 \qquad (61)$$

in cui:

$${}_s\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_s\hat{a}_1\, h_1(\hat{X}_i) + {}_s\hat{a}_2\, h_2(\hat{X}_i) + \ldots\ldots + {}_s\hat{a}_q\, h_q(\hat{X}_i) \qquad (62)$$

$${}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_{s-1}\hat{a}_1 h_1(\hat{X}_i) + {}_{s-1}\hat{a}_2\, h_2(\hat{X}_i) + \ldots\ldots + {}_{s-1}\hat{a}_q\, h_q(\hat{X}_i) \qquad (63)$$

rappresentano i modelli stimati rispettivamente alle iterazioni s ed s - 1.

Il procedimento richiede la conoscenza di almeno una stima iniziale dei parametri $a_1, a_2, \ldots, a_q$, che si ricava attraverso il metodo dei minimi quadrati basato sul vincolo (60).

Ottenuta tale stima, che indicheremo con ${}_o\hat{a}_1, {}_o\hat{a}_2, .., {}_o\hat{a}_q$,

si applica una seconda volta il metodo dei minimi quadrati minimizzando la funzione:

$$ {}_1S''(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i) - {}_1\hat{\hat{\epsilon}}^2(\hat{X}_i)}{{}_o\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right)^2 \qquad (64) $$

e ricavando le nuove stime ${}_1\hat{a}_1, {}_1\hat{a}_2, \ldots, {}_1\hat{a}_q$, essendo:

$$ {}_o\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_o\hat{a}_1\, h_1(\hat{X}_i) + {}_o\hat{a}_2\, h_2(\hat{X}_i) + \ldots + {}_o\hat{a}_q\, h_q(\hat{X}_i) \qquad (65) $$

$$ {}_1\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_1\hat{a}_1\, h_1(\hat{X}_i) + {}_1\hat{a}_2\, h_2(\hat{X}_i) + \ldots + {}_1\hat{a}_q\, h_q(\hat{X}_i) \qquad (66) $$

Il procedimento viene ripetuto fino a quando le stime ottenute non variano sostanzialmente da una iterazione all'altra (Cfr.nota XXII).

3.4 - Metodi per modelli non lineari nei parametri

Nei precedenti paragrafi abbiamo considerato il modello di regressione lineare nei parametri e di questo ne abbiamo dato la metodologia per la stima dei parametri incogniti.

Con la stessa metodologia e senza eccessive complicazioni formali si possono affrontare - specificando in modo diverso le funzioni $h_1(\hat{X}), h_2(\hat{X}), \ldots, h_q(\hat{X})$ nella (47) - i modelli illustrati nel paragrafo 2.1 :

a') $$ \hat{\epsilon}^2(\hat{X}) = a_1 + \frac{a_2}{\hat{X}} + u $$

b') $$ \log \hat{\epsilon}(\hat{X}) = a_1 + a_2 \log \hat{X} + u $$

c') $$ \hat{\epsilon}^2(\hat{R}) = \frac{a}{\hat{Y}} \frac{1 - \hat{R}}{\hat{R}} + u $$

d') $$ \hat{\epsilon}^2(\hat{R}) = a_1 + \frac{a_2}{\hat{R}\,\hat{Y}} + \frac{a_3}{\hat{Y}} + u $$

che sono tutte forme lineari nei parametri.

Per il secondo modello, ovviamente, ottenute da un campione le stime di a_1 e a_2 occorrerà risalire ai parametri del corrispondente modello non lineare (Cfr.nota XIII):

$$\hat{\epsilon}(\hat{X}) = \tilde{a}_1 \hat{X}^{\tilde{a}_2} \tilde{u} \qquad (67)$$

E' importante osservare che le proprietà delle stime di a_1 e a_2 non riproducono quelle di $\tilde{a}_1$ e $\tilde{a}_2$. Così, ad esempio, lo stimatore $\hat{a}_1$ é uno stimatore corretto del parametro a_1 : é quindi naturale stimare $\tilde{a}_1$ per mezzo di $\hat{\tilde{a}}_1 = \text{antilog } \hat{a}_1$, ma ciò non significa affatto che $\hat{\tilde{a}}_1$ così ottenuto sia uno stimatore corretto di a_1 (la media di una funzione non lineare non é generalmente uguale alla stessa funzione non lineare della media).

Ciò nonostante il criterio di applicare il metodo dei minimi quadrati a funzioni linearizzate nei parametri viene pure spesso applicato dai ricercatori, sia per comodità di calcolo, sia perché non posseggono un valido programma di stima non lineare, e sia per ottenere un valore di primo tentativo per i parametri da utilizzarsi per una successiva ricerca non lineare.

D'altra parte, qualora si voglia affrontare lo studio di relazioni non lineari interpolando i dati originali con funzioni non lineari senza l'ausilio di trasformazioni, si hanno notevoli complicazioni sia per quanto riguarda la stima dei parametri sia per una corretta analisi statistica dei risultati ottenuti.

La ricerca del minimo della funzione:

$$S(\hat{a}_1, \hat{a}_2, \ldots\ldots, \hat{a}_q) = \sum_1^p{}_i \left(\hat{\epsilon}^2(\hat{X}_i) - f(\hat{X}_i, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q) \right)^2 \qquad (68)$$

non é infatti così semplice come nel caso di modelli lineari nei

parametri, il problema si riduceva alla risoluzione di un sistema di ni lineari.

Si infatti vari ordini di difficoltà:

a) anzitutt rivazione analitica della funzione (68) rispetto ai pa può essere molto onerosa se non impossibile;

b) in secondo le condizioni che si ricavano, imponendo la uguagliazero di tali derivate, sono solo necessarie e non anche sufti, come nel caso lineare, ed é quindi possibile trovare tazionari che non rendono minima la funzione (68);

c) infine, luzione del sistema non lineare, che si ottiene imponendaglianza a zero delle derivate, può essere molto complessa impossibile, anche per semplici funzioni non lineari (a XXIII).

esti motivi il problema di stima non lineare é molto compleon sono finora disponibili procedimenti che forniscono la ne chiusa nel caso generale.

Svece disponibili delle filosofie abbastanza generali che ono di formulare i problemi di stima non lineari e di indie operazioni che occorrerebbe effettuare per ottenere la soe.

Ta, a causa delle difficoltà analitiche che tali operazioni ano, é soltanto introducendo ipotesi semplicative o per viassimata che si riesce a calcolare la soluzione del problema senso é molto frequente l'impiego di algoritmi iterativi colo per pervenire a determinazioni approssimate della sol

Qui di seguito illustreremo il metodo di linearizzazione iterativo, che richiede la conoscenza di una soluzione approssimata del problema o almeno di una stima iniziale per ricondurre il problema non lineare a un problema lineare approssimato, risolubile mediante procedimenti iterativi.

Per la complessità della sua trattazione formale ne daremo conto in termini prevalentemente euristici, affrontando l'argomento facendo riferimento al modello non lineare:

$$\hat{\hat{\epsilon}}^2(\hat{X}) = f(\hat{X}, \hat{a}_1, \hat{a}_2) = \hat{a}_1 \hat{X}^{\hat{a}_2} \tag{69}$$

in cui $\hat{a}_1$ e $\hat{a}_2$ sono valori approssimati dei parametri veri a_1 e a_2 .

Una regressione su questo modello suppone la minimizzazione di:

$$S(\hat{a}_1, \hat{a}_2) = \sum_1^p{}_i \left(\hat{\epsilon}^2(\hat{X}_i) - \hat{a}_1 \hat{X}_i^{\hat{a}_2} \right)^2 \tag{70}$$

dove $\hat{\epsilon}^2(\hat{X}_i)$ rappresenta il valore calcolato della varianza relativa corrispondente alla stima $\hat{X}_i$.

Supponiamo, per il momento, di conoscere i parametri veri a_1 e a_2 ed approssimiamo la $f(\hat{X}, \hat{a}_1, \hat{a}_2)$ con una funzione lineare intorno all'insieme iniziale costituito dai parametri veri.

A tal fine, linearizzando mediante la formula di Taylor e trascurando i termini di grado superiore al primo, si ottiene:

$$\hat{\hat{\epsilon}}^2(\hat{X}) = f(\hat{X}, \hat{a}_1, \hat{a}_2) \doteq f(\hat{X}, \hat{a}_1, \hat{a}_2)\Big|_{\substack{\hat{a}_1 = a_1 \\ \hat{a}_2 = a_2}} +$$

$$+ \frac{\partial f(\hat{X}, \hat{a}_1, \hat{a}_2)}{\partial \hat{a}_1}\Bigg|_{\substack{\hat{a}_1 = a_1 \\ \hat{a}_2 = a_2}} (\hat{a}_1 - a_1) + \frac{\partial f(\hat{X}, \hat{a}_1, \hat{a}_2)}{\partial \hat{a}_2}\Bigg|_{\substack{\hat{a}_1 = a_1 \\ \hat{a}_2 = a_2}} (\hat{a}_2 - a_2) \tag{71}$$

Avendosi:

$$\left.\frac{\partial f(\hat{X},\hat{a}_1,\hat{a}_2)}{\partial \hat{a}_1}\right|_{\substack{\hat{a}_1=a_1\\ \hat{a}_2=a_2}} = \hat{X}^{a_2} \quad e \quad \left.\frac{\partial f(\hat{X},\hat{a}_1,\hat{a}_2)}{\partial \hat{a}_2}\right|_{\substack{\hat{a}_1=a_1\\ \hat{a}_2=a_2}} = a_1\hat{X}^{a_2}\log \hat{X}$$

la (71) diviene:

$$\hat{\in}^2(\hat{Y}) \doteq a_1\hat{X}^{a_2} + \hat{X}^{a_2}(\hat{a}_1 - a_1) + a_1\hat{X}^{a_2}\log \hat{X}\ (\hat{a}_2 - a_2) \qquad (72)$$

Semplificando e riordinando, si ha la forma lineare nei parametri $\hat{a}_1$ e $\hat{a}_2$:

$$\hat{\in}^2(\hat{X}) \doteq \hat{a}_1(\hat{X}^{a_2}) + \hat{a}_2(a_1\hat{X}^{a_2}\log \hat{X}) - a_1a_2\hat{X}^{a_2}\log \hat{X} \qquad (73)$$

La minimizzazione di $S(\hat{a}_1,\hat{a}_2)$, definita dalla (70), é allora sostituita da quella di :

$$S(\hat{a}_1,\hat{a}_2) = \qquad (74)$$

$$= \sum_1^p{}_i \left[\hat{\in}^2(\hat{X}_i) - \left(\hat{a}_1(\hat{X}_i^{a_2}) + \hat{a}_2(a_1\hat{X}_i^{a_2}\log \hat{X}_i) - a_1a_2\hat{X}_i^{a_2}\log \hat{X}_i \right) \right]^2$$

Il minimo di questa funzione viene determinato annullando le sue derivate parziali: la soluzione del sistema lineare fornirebbe gli stimatori $\hat{a}_1$ e $\hat{a}_2$.

In realtà, e come é ovvio, non si conoscono i valori dei parametri veri a_1 e a_2 ; in pratica, per i calcoli di stima, si dovrà procedere ad iterazioni. Si comincia con l'attribuire ai parametri incogniti un determinato valore che indicheremo con i simboli ${}_o\hat{a}_1$ e ${}_o\hat{a}_2$; in tal modo é possibile stimare i parametri minimizzando la seguente espressione:

$$S(\hat{a}_1,\hat{a}_2) = \qquad (75)$$

$$= \sum_1^p{}_i \left[\hat{\in}^2(\hat{X}_i) - \left(\hat{a}_1(\hat{X}_i{}^{o\hat{a}_2}) + \hat{a}_2({}_o\hat{a}_1\hat{X}_i{}^{o\hat{a}_2}\log \hat{X}_i) - {}_o\hat{a}_1\,{}_o\hat{a}_2\,\hat{X}_i{}^{o\hat{a}_2}\log \hat{X}_i \right) \right]^2$$

Le stime così ottenute, che indichiamo con i simboli ${}_1\hat{a}_1$ e ${}_1\hat{a}_2$, vengono utilizzate come valori da attribuire ai para

metri incogniti a_1 e a_2 ;quindi si minimizza l'espressione:

$$S(\hat{a}_1, \hat{a}_2) = \tag{76}$$

$$= \sum_{1}^{p}{}_i \left[\hat{\epsilon}^2(\hat{X}_i) - \left(\hat{a}_1 (\hat{X}_i {}_1\hat{a}_2) + \hat{a}_2 ({}_1\hat{a}_1 \hat{X}_i {}_1\hat{a}_2 \log \hat{X}_i) - {}_1\hat{a}_1 \, {}_1\hat{a}_2 \, \hat{X}_i {}_1\hat{a}_2 \log \hat{X}_i \right) \right]^2$$

e si stimano i nuovi parametri ${}_2\hat{a}_1$ e ${}_2\hat{a}_2$ .Il procedimento di linearizzazione e stima viene ripetuto fino a quando le stime ottenute non variano sostanzialmente da un'iterazione all'altra. Se non c'é convergenza,le iterazioni dovranno prendere inizio da diversi valori arbitrari.

Prima di concludere questo paragrafo riteniamo utile osservare che,accanto al procedimento sopra descritto,esistono altri metodi di stima non lineare tra i quali ricordiamo il metodo di ottimizzazione diretta(una sua variante va sotto il nome di steepest descent) e il metodo di ricerca diretta del minimo (Cfr. [4] ,[20],[21],[23] ,[24] ,[25]).

4. Misura del grado di accostamento

L'ultimo stadio del procedimento di interpolazione consiste nella verifica dell'accostamento ottenuto,mediante la funzione interpolante,tra dati osservati e dati teorici,cioé nella misurazione del grado di approssimazione raggiunto surrogando i valori di $\hat{\epsilon}^2(\hat{X}_i)$ con i valori $\hat{\hat{\epsilon}}^2(\hat{X}_i)$,con i= 1,.....,p.

Dal punto di vista statistico un giudizio sulla bontà del modello di regressione può ottenersi già dal semplice esame grafico,ma é sempre opportuno ricorrere ad un indice sintetico.

La preferenza viene generalmente data all'indice di

determinazione, indicato con R^2, che può essere definito da (Cfr. nota XXIV):

$$R^2 = \frac{\text{Dev}\left[\hat{\hat{\in}}^2(\hat{X})\right]}{\text{Dev}\left[\hat{\in}^2(\hat{X})\right]} = \frac{\sum_1^p{}_i \left(\hat{\hat{\in}}^2(\hat{X}_i) - \bar{\hat{\in}}^2(\hat{X}_i) \right)^2}{\sum_1^p{}_i \left(\hat{\in}^2(\hat{X}_i) - \bar{\hat{\in}}^2(\hat{X}_i) \right)^2} \qquad (77)$$

Il significato della (77) é chiaro: esprime quanta parte della variabilità totale é spiegata dalla funzione di regressione. I valori di R^2 devono avvicinarsi quanto più possibile all'unità: valori inferiori a circa 0,9 indicano, generalmente, una cattiva rappresentazione del fenomeno da parte del modello.

5. Utilizzazione del metodo dei modelli regressivi nelle indagini reali

Nei precedenti paragrafi abbiamo analizzato, con un certo dettaglio, le forme funzionali di uso più frequente nella pratica della presentazione degli errori campionari mediante modelli; di queste ne abbiamo illustrato alcuni metodi di stima dei parametri e costruito degli indici per valutarne il grado di accostamento ai valori osservati.

La trattazione é stata sviluppata con riferimento ad un ipotetico gruppo G_s costituito dalle s stime $\hat{X}_1, \ldots, \hat{X}_i, \ldots, \hat{X}_s$, le quali devono essere caratterizzate da un deff costante o approssimativamente tale, affinché l'assunzione dei modelli discussi nei precedenti paragrafi possa essere giustificata.

Le considerazioni svolte, anche se ben formalizzate in ogni loro aspetto, non consentono tuttavia di enucleare una strategia finalizzata a contesti reali generalmente più complicati di quello cui le considerazioni svolte si riferiscono.

In questo paragrafo, quindi, presenteremo un approccio globale in base al quale é possibile operare in modo adeguato

nel caso di indagini effettive.

A tal fine supponiamo che una data indagine abbia fornito S stime di frequenze assolute, indicate con:

$$G_S = \left(\hat{X}_1, \ldots\ldots, \hat{X}_i, \ldots\ldots, \hat{X}_S\right)$$

Uno schema logico dell'approccio che intendiamo presentare é schematizzato nella figura 3.

Come si può notare dalla figura, gli stadi fondamentali sono sei.

Nel primo stadio si suddivide il gruppo G_S nei T sottogruppi:

$$G_{S_1} = \left(\hat{X}_{11}, \ldots\ldots, \hat{X}_{1i}, \ldots\ldots, \hat{X}_{1S_1}\right)$$

$$\vdots$$

$$G_{S_t} = \left(\hat{X}_{t1}, \ldots\ldots, \hat{X}_{ti}, \ldots\ldots, \hat{X}_{tS_t}\right)$$

$$\vdots$$

$$G_{S_T} = \left(\hat{X}_{T1}, \ldots\ldots, \hat{X}_{Ti}, \ldots\ldots, \hat{X}_{TS_T}\right)$$

sotto il vincolo che le stime appartenenti a ciascun sottogruppo abbiano un deff costante o approssimativamente tale. Questa condizione, anche se ad una prima impressione appare ragionevole, non é tuttavia realizzabile sul piano concreto. Infatti, per realizzare la suddivisione in oggetto nel rispetto della suddetta condizione, occorrerebbe calcolare, per ciascuna delle S stime, il deff che a sua volta comporta la determinazione dell'errore campionario effettivo (relativo cioé al campione usato per l'effettuazione dell'in

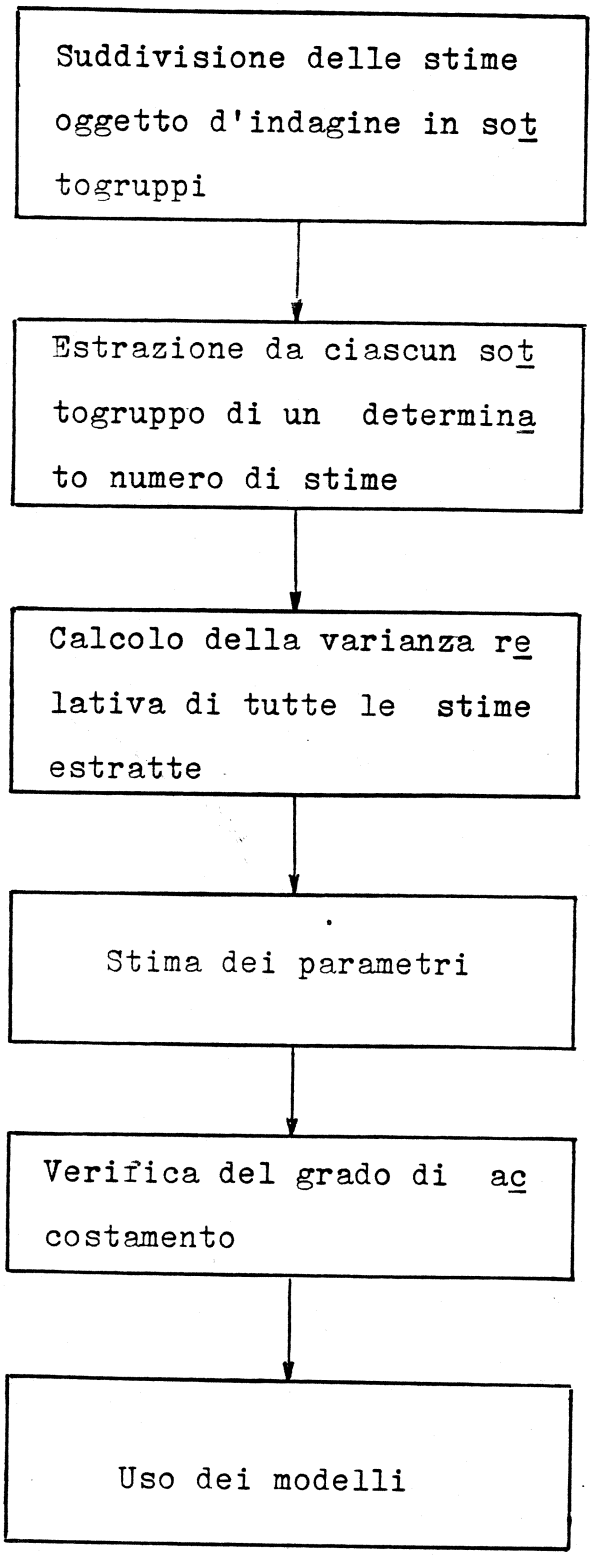
Suddivisione delle stime oggetto d'indagine in sottogruppi
Estrazione da ciascun sottogruppo di un determinato numero di stime
Calcolo della varianza relativa di tutte le stime estratte
Stima dei parametri
Verifica del grado di accostamento
Uso dei modelli

Fig. 3

dagine) e di quello corrispondente ad un ipotetico campione casuale semplice di pari numerosità,e ciò non é proponibile , in quanto comporterebbe un elevato numero di elaborazioni meccanografiche e di conseguenza tempi di calcolo e costi elevati.

In pratica,ogni ricercatore nell'affrontare la suddivisione in esame tiene conto delle esperienze passate e di alcune indicazioni di massima utili allo scopo,quali il riferire le stime sia al medesimo livello geografico che alle stesse caratteristiche demografiche ed economiche.

Una volta definita la suddivisione si passa (secondo stadio) alla scelta,nell'ambito di ciascun sottogruppo G_{S_t} ,di un campione di p_t stime,generalmente di dimensione piccola rispetto a quella del gruppo G_{S_t} .Questa situazione é riportata nel quadro seguente:

$$
\begin{aligned}
G_{p_1} &= \left(\hat{X}_{11}, \ldots\ldots, \hat{X}_{1i}, \ldots\ldots, \hat{X}_{1p_1} \right) \\
&\cdots \quad \cdots \quad \cdots \quad \cdots \\
G_{p_t} &= \left(\hat{X}_{t1}, \ldots\ldots, \hat{X}_{ti}, \ldots\ldots, \hat{X}_{tp_t} \right) \\
&\cdots \quad \cdots \quad \cdots \quad \cdots \\
G_{p_T} &= \left(\hat{X}_{T1}, \ldots\ldots, \hat{X}_{Ti}, \ldots\ldots, \hat{X}_{Tp_T} \right)
\end{aligned}
$$

La scelta delle stime $\hat{X}_{ti}$ é uno dei compiti più importanti della " programmazione della sperimentazione ".In questa fase ci si deve assicurare che il campo in cui il modello di regressione dovrà essere utilizzato sia sufficientemente esplorato o,come si suole dire,sia rappresentativo.

Nei casi in cui si abbia una sola variabile indipendente, come quelli trattati in questo studio, é abbastanza semplice assicurarsi che l'intervallo di validità del modello venga analizzato in modo corretto: a tal fine basta, ad esempio, scegliere stime approssimativamente equispaziate che ricoprano il campo sperimentale nel miglior modo possibile. Questa situazione é mostrata nella figura 4 con riferimento al generico gruppo G_{s_t}, il cui campo di validità del modello di regressione - avendo supposto senza alcuna perdita di generalità $X_{ti} \leq X_{t,i+1}$ per $i = 1, \ldots, S_t$ - risulta definito dall'intervallo $\left[X_{t1}, X_{ts_t}\right]$.

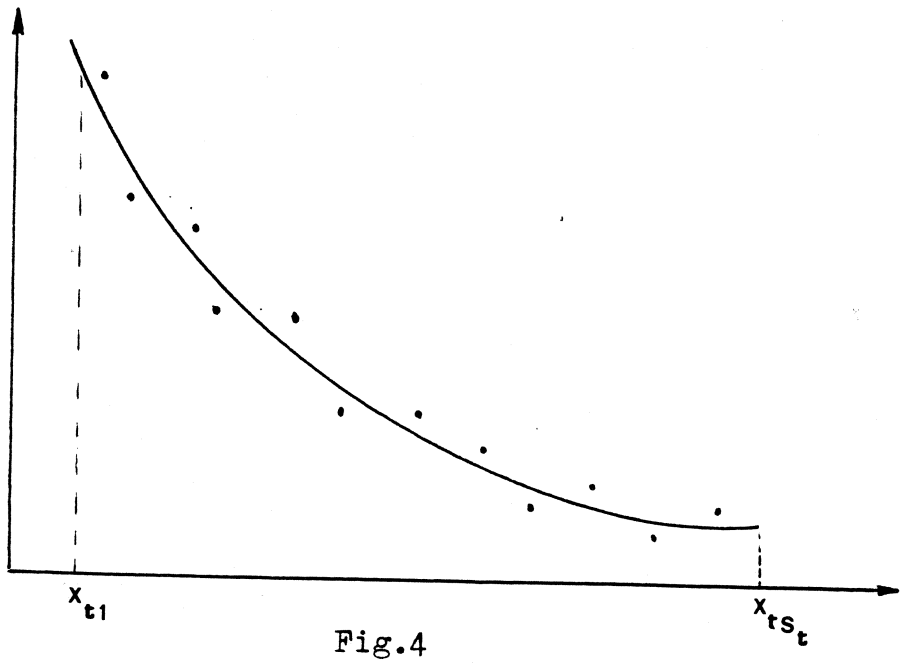

Fig.4

Nella fase seguente (terzo stadio) si determinano le varianze relative, $\hat{\epsilon}^2(\hat{X})$, di ciascuna delle stime appartenenti ai sottoinsiemi G_{p_t}, $t = 1, \ldots, T$.

Successivamente (quarto stadio), sulla base dei valori campionari $\left(\hat{X}_{ti}, \hat{\epsilon}^2(\hat{X}_{ti})\right)$, $i = 1, \ldots, p_t$, si ricavano per ogni gruppo le stime dei parametri del modello:

$$\hat{\epsilon}^2(\hat{X}) = a_1 + \frac{a_2}{\hat{X}} + u$$

oppure del modello alternativo:

$$\log \hat{\epsilon}(\hat{X}) = a_1 + a_2 \log \hat{X} + u$$

Una volta stimato il modello si passa (quinto stadio) alla valutazione del grado di accostamento, che avviene con il criterio già esposto.

Se il modello viene ritenuto idoneo (ed é questo l'ultimo stadio) si passa alla sua utilizzazione e nulla di specifico va aggiunto a quanto già detto nel paragrafo 2.1. Se il modello risulta non adeguato é necessario ricercare i motivi e la fonte di tale inadeguatezza e riformularlo.

6 . Considerazioni finali e prospettive di ricerca

Le considerazioni fin qui svolte sono servite ad illustrare una propost metodologica che può costituire un valido punto di riferimento per chiunque voglia accostarsi alla problematica della presentazione degli errori campionari nell' ottica dei modelli gressivi.

La nostrproposta,ovviamente,non voleva né poteva toccare ogni argomto,per cui,se da un lato abbiamo insistito sugli aspetti metodologici,interpretativi ed operativi ,nello stesso tempo abbiamo anche escluso argomenti specifici(ad esempio,tests per la verifica del modello stimato) e le applicazioni ad indagini ripetute.

Inoltre l'approccio illustrato in questa nota,pur trovando un crescente interesse soprattutto fra gli utilizzatori, non esaurisce il discorso sull'intera problematica della presentazione degli errori di campionamento mediante modelli.

Infatti esiste un altro,basato sull'effetto complessivo del disegno campionamento e su alcune sue componenti,che sta acquistando certo credito per merito di un ristretto numero di studiosi con vari saggi (Cfr. [44] ,[45]) hanno contribuito a chiarire le caratteristiche specifiche e le potenzialità analitiche. A parere nostro,tuttavia,tale approccio non ha ancora raggiunto una sistemazione definitiva e statisticamente valida;citiamo ad esempio l'impiego,nel caso di campioni non autoponderanti,dell'effetto ponderazione,la cui procedura di stima é ben lungi dal fornire risultati statisticamente corretti ed operativamente soddisfacenti.

In futuro dunque ci sembra che la ricerca dovrebbe essere orientata su tre diverse linee:

a) approfondimento teorico di alcuni aspetti metodologici non ancora completamente chiariti;

b) organizzazione dei due approcci metodologici in modo da disporre di un'analisi comparativa per individuarne portata e limiti;

c) applicazione ad indagini reali.

Note

I Nella fase preparatoria di ogni indagine, campionaria o totale, vari tipi di errore possono anzitutto scaturire da una inadeguata edificazione dei capisaldi del piano di rilevazione: e cioé da concetti mal precisati, da definizioni formulate in modo ambiguo o scarsamente comprensibili, da cattiva formulazione del modello di rilevazione e da inadeguatezza delle istruzioni predisposte per la sua compilazione. Anche gli elenchi-base possono divenire fonte di errori. Infatti, un elenco-base può essere: a) incompleto, se non contiene tutte le unità che definiscono l'universo; b) sovrabbondante, se contiene unità che non fanno parte dell'universo; c) non aggiornato, se non rispecchia la composizione strutturale e la situazione reale ad una data quanto più possibile prossima a quella della rilevazione; di conseguenza essi possono non rappresentare una base valida per la formazione di campioni rappresentativi.

Di grande interesse sono anche gli errori che sorgono per effetto delle mancate risposte, le quali si manifestano sia come mancate interviste, sia come assenza di risposta valida ad una domanda del modello di rilevazione.

Altre fonti di errore sono connesse alla personalità ed agli atteggiamenti del rilevatore e dell'intervistato ed alla loro interazione.

Una trattazione estesa sulle fonti di errore può essere trovata in [6], [11], [17], [39], [51].

II Generalmente in sede di pianificazione di ogni rilevazione campionaria, al fine di aumentare la precisione delle stime oggetto d'indagine, ci si sforza di ridurre l'errore campionario mediante l'impiego di schemi probabilistici e tecniche tra le quali la stratificazione, il campionamento con probabilità variabile, la post-stratificazione, le procedure di stima basate sul metodo del rapporto e della regressione (Cfr. [36]).

III La cattiva riuscita di una indagine é in genere dovuta più agli errori di risposta (o di misura) che a quelli di campionamento. Di conseguenza, nel predisporre una rilevazione statistica, lo sforzo viene essenzialmente indirizzato nella costruzione di un sistema coerente di concetti ed operazioni e nella messa in moto di validi sistemi di controllo che consentano di ridurre, già durante le fasi immediatamente precedenti la rilevazione o durante il suo svolgimento, l'azione deformante dei fattori connessi a difetti organizzativi. Grande importanza assumono, nella fase organizzativa, l'identificazione dei concetti-base, la formulazione delle necessarie definizioni, la redazione del modello di rilevazione, la ricerca della migliore caratterizzazione del campo di osservazione, la predisposizione di istruzioni idonee ed esemplificate, la realizzazione di un buon addestramento dei rilevatori sulla base di un programma da condurre con criteri uniformi. A parte queste misure squisitamente organizzative, vengono inoltre frequentemente utilizzate alcune tec

niche che consentono di combattere efficacemente la deformazione nelle risposte, come ad esempio la tecnica basata sulle classi di ponderazioni che consente di ridurre gli effetti distorsivi dovuti alle mancate risposte.

IV L'affidabilità di una stima campionaria é l'unione di due componenti: la precisione (o efficienza), data dal reciproco della varianza della stima, e l'accuratezza, che riguarda tutti gli errori di risposta (o di misura) cumulati in ogni fase di formazione dei dati.

V Le statistiche ausiliarie di cui é stata fatta menzione nell'Introduzione sono l'effetto complessivo del disegno di campionamento o deff e le sue principali componenti (effetti stratificazione, clustering e ponderazione) e il coefficiente di correlazione intra-classe. Tali statistiche, oltre a descrivere in modo più approfondito la natura delle variabili oggetto d'indagine, sono molto utili:

- nella valutazione dei piani di campionamento utilizzati per l'effettuazione di indagini concrete (Cfr. [7], [8], [22], [28], [33], [35], [36], [44]);
- nella predisposizione di future indagini (Cfr. [7], [8], [17], [19]);
- nella costruzione di modelli per la presentazione degli errori di campionamento (Cfr. [31], [32], [38], [45], [46]);
- nell'inferenza statistica (Cfr. [9], [10], [15], [18], [26]).

VI In occasione dell'indagine campionaria sulle vacanze e gli sports, eseguita dall'ISTAT nel mese di novembre 1985, é stata presa l'iniziativa di incorporare nel piano di rilevazione il controllo della varianza correlata di risposta che, come dimostrano i molti esempi illustrati nella letteratura disponibile, rappresenta la componente dominante della varianza di risposta (Cfr. [7], [14], [42], [43]).

VII Per l'effettuazione di indagini campionarie su larga scala si ricorre generalmente all'impiego di disegni di campionamento che nella letteratura statistica sull'argomento sono definiti " complessi ". Nella maggior parte dei casi, infatti, si tratta di disegni a più stadi di campionamento, stratificati, con selezione delle unità primarie con probabilità variabile e procedimenti di stima basati sul metodo del rapporto e della regressione. Inoltre, si tratta di indagini che hanno la finalità di fornire diversi tipi di stime (frequenze assolute e relative, medie, rapporti, totali), per ciascuno dei quali bisogna determinare un numero elevato di parametri. (Cfr. [22], [32], [34], [38], [46], [49]).

VII Fra i vari metodi suggeriti per l'ottenimento di una valutazione approssimata della varianza di campionamento citiamo quelli più soddisfacenti e maggiormente usati:

- metodo dello sviluppo in serie di Taylor (o di linearizzazione);
- metodo delle replicazioni ripetute bilanciate;
- metodo di jack-knife;
- metodo dei gruppi casuali.

Per una trattazione approfondita di tali metodi si ve da Wolter [46]. Segnaliamo inoltre [8], [13], [16], [19], [22], [28], [29], [30], [31], [34], [38], [44], [47], [49].

IX Tali modelli si possono raggruppare in due tipi, distinti dalla diversa tecnica di approccio: quella dei modelli regressivi e quella basata sull' effetto del disegno di campionamento (Cfr. [1], [22], [32], [37], [44], [45], [46], [50]).

X Così, ad esempio, per l'indagine ISTAT sulle condizioni di salute della popolazione italiana nel 1980, relativamente al caso delle stime delle frequenze assolute, i valori delle $\hat{X}_i'$ sono stati fissati pari a : 10, 20, 40, 60, 80, 100, 200, 400, 600, 800, 1.000, 2.000, 4.000, 6.000, 8.000, 10.000, 20.000, 30.000, 40.000 e 50.000 migliaia (Cfr. [49]).

XI Accanto ai modelli suddetti sono stati suggeriti altri modelli di uso meno frequente (Cfr. [5]).

XII L'effetto del disegno di campionamento in sostanza misura l'efficienza del disegno complesso rispetto al disegno casuale semplice, naturalmente di uguale numerosità in termini di unità finali (Cfr. [7], [8], [17], [36]).

XIII Il corrispondente modello non lineare é definito dalla forma:

$$\in (\hat{X}) = \tilde{a}_1 \, X^{\tilde{a}_2} \, \tilde{u} \qquad (1')$$

Applicando infatti l'operazione di logaritmo

ad ambo i membri della (1') si ottiene:

$$\log \in(\hat{Y}) = \log \tilde{a}_1 + \tilde{a}_2 \log X + \log \tilde{u} \qquad (2')$$

e ponendo:

$$\log \tilde{a}_1 = a_1 \; ; \; \tilde{a}_2 = a_2 ; \; \log \tilde{u} = u \qquad (3')$$

la (2') assume la forma:

$$\log \in(\hat{X}) = a_1 + a_2 \log X + u \qquad (4')$$

formalmente coincidente con la (18).

XIV Ci sembra utile illustrare alcune possibili utilizzazioni delle relazioni (32) e (34), che serviranno anche a meglio chiarire i concetti sinora esposti.

A tal fine consideriamo in primo luogo le relazioni operative corrispondenti della (32) e (34), definite dalle espressioni:

$$\hat{\in}^2(\hat{R}) = \hat{\in}^2(\hat{X}) - \hat{\in}^2(\hat{Y}) \qquad (5')$$

$$\hat{\in}^2(\hat{R}) = \frac{a_2}{\hat{Y}} \frac{1 - \hat{R}}{\hat{R}} + u \qquad (6')$$

dove $\hat{\in}^2(\hat{R})$, $\hat{\in}^2(\hat{X})$, $\hat{\in}^2(\hat{Y})$, $\hat{X}$, $\hat{Y}$ ed $\hat{R}$ indicano rispettivamente le stime di $\in^2(\hat{R})$, $\in^2(\hat{X})$, $\in^2(\hat{Y})$, X ,Y ed R.

Supponiamo ora che si voglia determinare una stima della varianza relativa di $\hat{R} = \hat{X} / \hat{Y}$, in cui $\hat{X}$ é un sottoinsieme di $\hat{Y}$.

Dobbiamo a questo punto distinguere i due casi seguenti:

a) - Le stime $\hat{X}$ ed $\hat{Y}$ appartengono rispettivamente ai due seguenti gruppi:

$$G_s(\hat{X}) = \left(\hat{X}_1, \ldots\ldots, \hat{X}_i, \ldots\ldots, \hat{X}_s \right)$$

$$G_{\bar{s}}(\hat{Y}) = \left(\hat{Y}_1, \ldots\ldots, \hat{Y}_i, \ldots\ldots, \hat{Y}_{\bar{s}} \right)$$

Siano inoltre:

$$\hat{\hat{\epsilon}}^2(\hat{X}_i) = \hat{a}_1' + \frac{\hat{a}_2'}{\hat{X}_i} \qquad (7')$$

$$\hat{\hat{\epsilon}}^2(\hat{Y}_i) = \hat{a}_1'' + \frac{\hat{a}_2''}{\hat{Y}_i} \qquad (8')$$

i modelli stimati, mediante i quali é possibile determinare una stima della varianza relativa di ciascuna delle stime appartenenti ai due gruppi $G_s(\hat{X})$ e $G_{\bar{s}}(\hat{Y})$.

Pertanto, sostituendo nella (5') i valori calcolati delle varianze relative $\hat{\epsilon}^2(\hat{X})$ e $\hat{\epsilon}^2(\hat{Y})$ con i corrispondenti valori teorici espressi dai modelli (7') e (8'), si ottiene una stima della varianza relativa di $\hat{R}$ data da:

$$\hat{\hat{\epsilon}}^2(\hat{R}) = \hat{\hat{\epsilon}}^2(\hat{X}) - \hat{\hat{\epsilon}}^2(\hat{Y}) = \qquad (9')$$

$$= \hat{a}_1' + \frac{\hat{a}_2'}{\hat{X}} - \hat{a}_1'' - \frac{\hat{a}_2''}{\hat{Y}}$$

Una volta determinata la varianza relativa, possono ricavarsi l'errore relativo e l'errore assoluto.

Allo scopo di rendere più agevole il calcolo di $\hat{\hat{\epsilon}}^2(\hat{R})$, nelle relazioni finali sui risultati delle inda

gini effettuate sono presentati dei prospetti - simili a a quello descritto a pagina 7 del paragrafo 2.1 - sulla base dei quali é possibile pervenire ad una stima di $\epsilon^2(\hat{R})$.

Più precisamente, da ciascuno di tali prospetti - uno per ogni gruppo - con riferimento alle due stime $\hat{X}$ ed $\hat{Y}$, si ricavano i corrispondenti errori relativi $\hat{\hat{\epsilon}}(\hat{X})$ e $\hat{\hat{\epsilon}}(\hat{Y})$; la differenza dei loro quadrati fornisce poi la stima $\hat{\hat{\epsilon}}^2(\hat{R})$.

b)- Le stime $\hat{X}$ ed $\hat{Y}$ appartengono allo stesso gruppo:

$$G_s = \left(\hat{X}_1, \ldots\ldots, \hat{X}_i = \hat{X}, \ldots\ldots, \hat{X}_j = \hat{Y}, \ldots, \hat{X}_s \right)$$

per il quale il modello stimato é definito dalla forma:

$$\hat{\hat{\epsilon}}^2(\hat{X}) = \hat{a}_1 + \frac{\hat{a}_2}{\hat{X}} \qquad (10')$$

Allora una stima della varianza relativa, espressa dalla (6'), può ottenersi attraverso la funzione:

$$\hat{\hat{\epsilon}}^2(\hat{R}) = \frac{\hat{a}_2}{\hat{Y}} \frac{I - \hat{R}}{\hat{R}} \qquad (11')$$

Prima di concludere questa nota presentiamo un'ultima applicazione, basata sull 'uso del modello (6'), che generalizza i risultati di cui al precedente punto b).

Dato il gruppo:

$$G_s = \left(\hat{X}_1, \ldots\ldots\ldots, \hat{X}_i, \ldots\ldots\ldots, \hat{X}_s \right)$$

supponiamo che il modello:

$$\hat{\epsilon}^2(\hat{X}) = a_1 + \frac{a_2}{\hat{X}} + u \qquad (12')$$

sia teoricamente valido a descrivere la relazione tra la

varianza relativa e la stima.

Sia poi:

$$G_p = \left(\hat{X}_1, \ldots\ldots, \hat{X}_i, \ldots\ldots, \hat{X}_p\right)$$

un sottoinsieme costituito da p stime estratte dall'insieme G_s ed indichiamo con:

$$\hat{\epsilon}^2(\hat{X}_1), \ldots\ldots, \quad \hat{\epsilon}^2(\hat{X}_i), \ldots\ldots, \quad \hat{\epsilon}^2(\hat{X}_p)$$

le corrispondenti stime delle varianze relative.

Tramite il metodo dei minimi quadrati si determinano le stime $\hat{a}_1$ e $\hat{a}_2$ che sono funzioni esclusive dei valori $\left(\hat{\epsilon}^2(\hat{X}_i), \hat{X}_i\right)$, $i = 1, \ldots, p$; il modello stimato é pertanto espresso da:

$$\hat{\epsilon}^2(\hat{X}) = \hat{a}_1 + \frac{\hat{a}_2}{\hat{X}} + e \qquad (13')$$

da cui segue che una stima di $\hat{\epsilon}^2(\hat{X})$ può ottenersi mediante la forma:

$$\hat{\hat{\epsilon}}^2(\hat{X}) = \hat{a}_1 + \frac{\hat{a}_2}{\hat{X}} \qquad (14')$$

Ciò posto, consideriamo i seguenti diversi valori di $\hat{R}$ e di $\hat{Y}$:

$$\hat{R} = \left(\hat{R}_1, \ldots\ldots\ldots, \hat{R}_g, \ldots\ldots\ldots, \hat{R}_G\right)$$

$$\hat{Y} = \left(\hat{Y}_1, \ldots\ldots\ldots, \hat{Y}_j, \ldots\ldots\ldots, \hat{Y}_J\right)$$

Così, ad esempio, nell'indagine statunitense CPS (Current Population Survey) sono stati considerati i seguenti valori:

$$\hat{R} = \left(0{,}02;\ 0{,}05;\ 0{,}10;\ 0{,}25;\ 0{,}50;\ 0{,}75;\ 0{,}90;\ 0{,}95;0{,}98\right)$$

$\hat{Y} = (250.000;\ 500.000;\ 1.000.000;\ 2.500.000;\ 5.000.000;$
$10.000.000;\ 25.000.000;\ 50.000.000)$

Tenendo presente il modello (11'),ne segue che una stima della varianza relativa di $\hat{R}_g$,sotto la condizione che $\hat{Y}_j$ sia l'ampiezza della base di $\hat{R}_g$ .si ricava usando la relazione:

$$\hat{\hat{\epsilon}}^2(\hat{R}_g) = \frac{\hat{a}_2}{\hat{Y}_j} \quad \frac{I - \hat{R}_g}{\hat{R}_g} \qquad (15')$$

Facendo variare g e j si ottengono G • J stime di varianza relativa,dalle quali si ricavano subito G•J, errori relativi che,riportati in un prospetto del tipo sotto indicato,consentono di pervenire ad una valutazione del livello di precisione di qualsiasi stima ottenuta come rapporto di due stime campionarie $\hat{X}$ ed $\hat{Y}$ con $\hat{X} \subset \hat{Y}$.

	$\hat{R}_1$ ..... $\hat{R}_g$ ..... $\hat{R}_G$
$\hat{Y}_1$	$\hat{\hat{\epsilon}}_{11}$ ..... $\hat{\hat{\epsilon}}_{1g}$ ..... $\hat{\hat{\epsilon}}_{1G}$
$\hat{Y}_j$	$\hat{\hat{\epsilon}}_{j1}$ ..... $\hat{\hat{\epsilon}}_{jg}$ ..... $\hat{\hat{\epsilon}}_{jG}$
$\hat{Y}_J$	$\hat{\hat{\epsilon}}_{J1}$ ..... $\hat{\hat{\epsilon}}_{Jg}$ ..... $\hat{\hat{\epsilon}}_{JG}$

Così,ad esempio,se il rapporto tra la stima $\hat{X}$ (numero di fumatori maschi nella classe di età 30-40 anni) e la stima $\hat{Y}$ (numero di maschi nella classe di età 30-40) é pari a $\hat{R}_g$ e se $\hat{Y}$ é uguale a $\hat{Y}_j$,l'errore campionario relativo

della stima $\hat{X} / \hat{Y}$ é pertanto pari a $\hat{\hat{\epsilon}}_{jg}$, avendo posto:

$$\hat{\hat{\epsilon}}_{jg} = \hat{\hat{\epsilon}} \left(\hat{R}_g / \hat{Y}_j \right) \qquad (16')$$

Per la determinazione dell'errore relativo di una stima rapporto il cui valore non risulta compreso fra i valori costituenti l'insieme:

$$\hat{R} = (\hat{R}_1, \ldots\ldots, \hat{R}_g, \ldots\ldots, \hat{R}_G)$$

é preferibile ricorrere all'impiego della funzione (15') anziché al prospetto in oggetto la cui utilizzazione richiederebbe una interpolazione sia rispetto ad $\hat{R}$ che ad $\hat{Y}$.

Osserviamo infine che il modello stimato (15') può ottenersi anche interpolando un determinato insieme di punti relativi a valori diversi di $\hat{R}$ e di $\hat{Y}$.

XV In pratica i modelli (42) e (43) sono sostituiti dai corrispondenti modelli operativi:

$$\hat{\epsilon}^2(\hat{R}) = a_1 + \frac{a_2}{\hat{X}} + \frac{a_3}{\hat{Y}} + u \qquad (17')$$

$$\hat{\epsilon}^2(\hat{R}) = a_1 + \frac{a_2}{\hat{R}\,\hat{Y}} + \frac{a_3}{\hat{Y}} + u \qquad (18')$$

XVI Per la presentazione degli errori campionari si può seguire un procedimento analogo a quello illustrato al punto b) della precedente nota XIV.

XVII La condizione (45) é necessaria,ma in generale non é sufficiente.Per sapere se $\hat{a}_1,\hat{a}_2,....,\hat{a}_q$,é un punto di minimo per la $S(\hat{a}_1,\hat{a}_2,...,\hat{a}_q)$ é necessario calcolare le derivate parziali seconde ed esaminare il segno di una loro particolare funzione.

Tuttavia se $f(\hat{X},a_1,a_2,....,a_q)$ é della forma:

$$a_1h_1(\hat{X}) + a_2h_2(\hat{X}) + ...+ a_qh_q(\hat{X}) \qquad (19')$$

in cui $h_1(\hat{X}),h_2(\hat{X}),.....,h_q(\hat{X})$ sono funzioni matematiche della sola $\hat{X}$ non contenenti ulteriori parametri,cioé se la (19') é lineare nei parametri,le (45) si riducono ad un sistema di equazioni lineari.In questo caso se il determinante dei coefficienti del sistema é diverso da zero,e tale ipotesi é generalmente verificata in pratica, il sistema ammette una sola soluzione che si dimostra essere quella che fornisce il minimo per la $S(\hat{a}_1,\hat{a}_2,...,\hat{a}_q)$.

XVIII Nel senso che le stime dei parametri possono ottenersi attraverso espressioni che non contengono altri parametri.

XIX Le proprietà a),b) e c) sono note con il nome di Teorema di Gauss - Markov.

XX La denominazione di ponderate,attribuita alle stime $\hat{a}_1,\hat{a}_2,....,\hat{a}_q$,deriva dal fatto che esse sono tutte derivabili dalla minimizzazione espressa dalla (56) in cui le p_i rappresentano dei pesi.

XXI Sono stati elaborati dei tests che permettono di decidere se esiste o meno eteroschedasticità.Per una trattazione estesa su questo argomento si veda [23] .

XXII A chiarimento di quanto detto, nell'esempio seguente calcoliamo la stima dei parametri del modello:

$$\hat{\epsilon}^2(\hat{X}) = a_1 + \frac{a_2}{\hat{X}} + u \qquad (20')$$

Applicando la (61) al nostro problema si ha:

$${}_sS''(\hat{a}_1, \hat{a}_2) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i) - {}_s\hat{\hat{\epsilon}}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right)^2 \qquad (21')$$

in cui:

$${}_s\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_s\hat{a}_1 + \frac{{}_s\hat{a}_2}{\hat{X}_i} \qquad (22')$$

$${}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i) = {}_{s-1}\hat{a}_1 + \frac{{}_{s-1}\hat{a}_2}{\hat{X}_i} \qquad (23')$$

Sostituendo la (22') nella (21') si ottiene:

$${}_sS''(\hat{a}_1, \hat{a}_2) = \sum_1^p{}_i \left(\frac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \frac{{}_s\hat{a}_1}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \frac{{}_s\hat{a}_2}{\hat{X}_i \; {}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right)^2 \qquad (24')$$

Imponendo l'annullamento delle derivate parziali della (24') rispetto ai parametri ${}_s\hat{a}_1$ e ${}_s\hat{a}_2$ si ha:

$$\begin{cases} \dfrac{\partial {}_sS''(\hat{a}_1, \hat{a}_2)}{\partial \, {}_s\hat{a}_1} = 2\sum_1^p{}_i \left(\dfrac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \dfrac{{}_s\hat{a}_1}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \dfrac{{}_s\hat{a}_2}{\hat{X}_i \; {}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right) \dfrac{-1}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} = 0 \\ \qquad (25') \\ \dfrac{\partial {}_sS''(\hat{a}_1, \hat{a}_2)}{\partial \, {}_s\hat{a}_2} = 2\sum_1^p{}_i \left(\dfrac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \dfrac{{}_s\hat{a}_1}{{}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} - \dfrac{{}_s\hat{a}_2}{\hat{X}_i \; {}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} \right) \dfrac{-1}{\hat{X}_i \; {}_{s-1}\hat{\hat{\epsilon}}^2(\hat{X}_i)} = 0 \end{cases}$$

che, semplificando opportunamente, si riduce al sistema di due equazioni lineari nelle due incognite ${}_s\hat{a}_1$ e ${}_s\hat{a}_2$:

$$\begin{cases} \left(\sum_1^p{}_i \dfrac{1}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right){}_s\hat{a}_1 + \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right){}_s\hat{a}_2 = \sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)} \\ \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right){}_s\hat{a}_1 + \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i^2\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right){}_s\hat{a}_2 = \sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)} \end{cases} \qquad (26')$$

Le relazioni del sistema (26') sono le equazioni normali la cui soluzione é:

$$\begin{cases} {}_s\hat{a}_1 = \dfrac{\left(\sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i^2\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right) - \left(\sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)}{\left(\sum_1^p{}_i \dfrac{1}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i^2\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right) - \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)^2} \\ {}_s\hat{a}_2 = \dfrac{\left(\sum_1^p{}_i \dfrac{1}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right) - \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{\hat{\epsilon}^2(\hat{X}_i)}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)}{\left(\sum_1^p{}_i \dfrac{1}{{}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)\left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i^2\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right) - \left(\sum_1^p{}_i \dfrac{1}{\hat{X}_i\ {}_{s-1}\hat{\hat{\epsilon}}^4(\hat{X}_i)}\right)^2} \end{cases} \qquad (27')$$

Per avviare il processo di iterazione, come abbiamo sottolineato nel paragrafo 3.3, é necessario disporre di una stima iniziale dei parametri a_1 e a_2; tale stima, che indicheremo con ${}_o\hat{a}_1$

e ${}_{o}\hat{a}_2$,può ottenersi minimizzando l'espressione (60) oppure direttamente dalle (27') sostituendo $\hat{\in}^4(\hat{X}_i)$ al posto di ${}_{s-1}\hat{\hat{\in}}^4(\hat{X}_i)$.

Una volta determinate le stime ${}_{o}\hat{a}_1$ e ${}_{o}\hat{a}_2$,sempre mediante le (27'),si ricavano (prima iterazione,s = 1) le stime ${}_{1}\hat{a}_1$ e ${}_{1}\hat{a}_2$ .tenendo presente che in base alla (23') si ha:

$$ {}_{o}\hat{\hat{\in}}^4(\hat{X}_i) = \left({}_{o}\hat{a}_1 + \frac{{}_{o}\hat{a}_2}{\hat{X}_i} \right)^2 \qquad (28') $$

Si passa quindi alla seconda iterazione(s = 2) e si stimano i nuovi parametri ${}_{2}\hat{a}_1$ e ${}_{2}\hat{a}_2$,avendosi:

$$ {}_{1}\hat{\hat{\in}}^4(\hat{X}_i) = \left({}_{1}\hat{a}_1 + \frac{{}_{1}\hat{a}_2}{\hat{X}_i} \right)^2 \qquad (29') $$

Continuando ad iterare con i valori stimati dei parametri stessi,si ha la convergenza delle stime da quando ogni iterazione procura valori di tali stime più vicini ai valori dell'iterazione precedente che non questi ultimi ai valori dell'iterazione precedente ancora.L'ultima iterazione é quella i cui valori di stima non scartano dai valori dell'iterazione precedente oltre un intervallo prefissato;l'ultima iterazione procura le stime richieste.

Così,ad esempio,nel caso dell'indagine statunitense " National Health Interview Survey"le condizioni sono:

$$ \left| \frac{{}_{s}\hat{a}_1 - {}_{s-1}\hat{a}_1}{{}_{s}\hat{a}_1} \right| \leqslant 0{,}01 \qquad \text{e} \qquad \left| \frac{{}_{s}\hat{a}_2 - {}_{s-1}\hat{a}_2}{{}_{s}\hat{a}_2} \right| \leqslant 0{,}01 $$

XXIII Ad esempio, applicando i minimi quadrati al modello di regressione:

$$Y = a_1 a_2^X + u \tag{30'}$$

si ottiene il sistema di equazioni:

$$\begin{aligned} \sum Y_i a_2^{X_i} &= a_1 \sum a_2^{2X_i} \\ a_1 \sum Y_i X_i a_2^{X_i - 1} &= a_1^2 \sum X_i a_2^{2X_i - 1} \end{aligned} \tag{31'}$$

la cui risoluzione é molto complicata.

XXIV Nella (77) abbiamo posto:

$$\bar{\hat{\epsilon}}^2(\hat{X}_i) = \frac{1}{p} \sum_1^p{}_i \hat{\epsilon}^2(\hat{X}_i) \tag{32'}$$

Riteniamo inoltre utile osservare che alcuni autori preferiscono usare altri indici per misurare la bontà di una interpolazione; citiamo, ad esempio, gli indici basati sui residui relativi espressi formalmente dalle formule:

$$I_1 = \frac{1}{p} \sum_1^p{}_i \frac{\left| \bar{\hat{\epsilon}}^2(\hat{X}_i) - \hat{\epsilon}^2(\hat{X}_i) \right|}{\hat{\epsilon}^2(\hat{X}_i)} \tag{33'}$$

$$I_2 = \frac{1}{p} \sum_1^p{}_i \frac{\left(\bar{\hat{\epsilon}}^2(\hat{X}_i) - \hat{\epsilon}^2(\hat{X}_i) \right)^2}{\hat{\epsilon}^2(\hat{X}_i)} \tag{34'}$$

Bibliografia

[1] BEAN J.A)),Estimation and Sampling Variance in the Health Irw Survey,Vital and Health Statistics,Serie 2,N.38.

[2] BUREAU DESTIQUE DES NATIONS UNIES (I975),La préparation des ts sur les enquêtes par sondage,Etudes statistiques C,N.I,New York.

[3] DALENIUS 2),Recent advances in sample survey theory and methoals of Mathematical Statistics,Vol.33.

[4] DRAPER N.MITH H.(I967),Applied Regression Analysis, New York,y.

[5] EDELMAN M67),Curve Fitting of Keyfitz Variance,Unpublished mdum,U.S.Bureau of the Census,Washington,D.C. 20233.

[6] FABBRIS L),Metodi statistici per l'analisi e il controllo dealità dei dati sanitari,in P.Bellini,Rigatti Luchinan (a cura di),Statistica e ricerca epidemiologicaa,CIEUP.

[7] FABBRIS L),Alcune proposte sul tema del sovracampionamento su regionale e provinciale dell'indagine sulle forze oro,Economia e Lavoro,Anno XV,n.3.

[8] FABBRIS L.I),L'errore di campionamento nell'indagine sull'assis economica pubblica nel Veneto:calcolo e consideraaggiuntive,in Emarginazione come sviluppo, Padova,CLE

[9] FELLEGI I.P. (I980), Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples, Journal of the American Statistical Association, Vol.75.

[10] FULLER W.A. (I975), Regression Analysis for Sample Surveys, Sankhya, Serie C, Vol.37.

[11] GIUSTI F. (I969), Su gli errori di osservazione nei censimenti e nelle rilevazioni campionarie, Atti della XXVI Riunione della S.I.S., Vol.II, Firenze.

[12] GONZALES M., OGUS J.L. e TEPPING B.J. (I975), Standards for Discussion and Presentation of Errors in Survey and Census Data, Journal of the American Statistical Association, Vol.70, N.35I, Part.II.

[13] GURNEY M. (I969), Random Group Method for Estimating Variance, Unpublished manuscript U.S. Bureau of the Census, DC 20233, Washington.

[14] HANSEN M.H., HURWITZ W.N. e BERSHAD M.A. (I96I), Measurement errors in census and surveys, Bollettino dell'Istituto Internazionale di Statistica, Tomo XXXVIII, Tokyo.

[15] HOLT D., SCOTT A.J. e EWINGS P.O. (I980), Chi-squared Test with Survey Data, Journal of the Royal Statistical Society, Sec. A, Vol.43.

[16] KEYFITZ N. (I957), Estimates of Sampling Variance when two unit are selected from each stratum, Journal of the American Statistical Association, Vol.52.

[17] KISH L. (I965), Survey Sampling, Wiley, New York.

[18] KISH L. e FRANKEL M.R. (I974), Inference from Complex Samples, Journal of the Royal Statistical Society, B.36.

[19] KISH L., GROVES R.M. e KROTKI K.P. (I976), Sampling errors for fertility surveys, Occasional Papers, N.I7, World Fertility Survey.

[20] MALINVAUD E., (I97I), Metodi statistici dell'econometria UTET, Torino.

[21] MARQUARDT D.W. (I963), An Algorithm for Least Squares Estimation of Nonlinear Parameters, in SIAM Journal of Numerical Analysis, n.II, Vol.2.

[22] NAPOLITANO P., RUSSO A. e ZANNELLA F. (I983), Calcolo, presentazione ed analisi degli errori di campionamento nell'indagine ISTAT sulle condizioni di salute della popolazione italiana e sul ricorso ai servizi sanitari, Atti del Convegno della S.I.S., Trieste.

[23] PICCOLO D. e VITALE C. (I98I), Metodi statistici per la analisi economica, Il Mulino, Bologna.

[24] RALSTON A. (I965), A First Course in Numerical Analysis, Mc Graw Hill, New York.

[25] RALSTON A. e JENNRICH R.I. (I978), A Derivate - Free Algorithm for nonlinear Least Squares, Technometrica, N.I.

[26] RAO J.N.K. e HIDIROGLOU M.A. (I98I), Chi-squares for the analysis of categorical data from the Canada Health Survey, Bollettino dell'Istituto Internazionale di Statistica, Buenos Aires.

[27] RUSSO A. e DI TRAGLIA M. (I982), Distribuzione per età della popolazione scolastica, anno I978-79: Grado di attendibilità dei risultati, Supplemento al Bollettino di Statistica, N.25, ISTAT, Roma.

[28] RUSSO A. (I984), Calcolo ed analisi degli errori di campionamento nell'indagine ISTAT sulle vacanze e gli sports degli italiani, anno I983, Atti della XXXII Riunione della S.I.S., Sorrento.

[29] RUSSO A. e NAPOLITANO P. (I984), L'errore di campionamento nell'indagine ISTAT sulle forze di lavoro: calcolo e presentazione in due regioni italiane, Atti della XXXII Riunione della S.I.S., Sorrento.

[30] RUSSO A. (I984), Indagine sulle vacanze, i viaggi e gli sport degli italiani nel I982: Piano della rilevazione campionaria ed errori di campionamento, Supplemento al Bollettino mensile di statistica, n.I5, ISTAT, Roma.

[31] RUSSO A. e DI TRAGLIA M. (I985), Indagine sulle strutture ed i comportamenti familiari: Disegno di campionamento, calcolo e presentazione degli errori campionari, ISTAT, Roma.

[32] RUSSO A. (I985), Modelli per la presentazione degli errori standard in campioni complessi, Giornate di metodologia statistica, Bressanone.

[33] RUSSO A. (I985), Su un metodo di stima degli effetti stratificazione e clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio, Quaderni di Discussione N.5, ISTAT, Roma.

[34] RUSSO A. e FALORSI P.D.(I985),Rilevazioni campionarie delle forze di lavoro:Metodologia del campionamento,calcolo e presentazione errori campionari,Quaderni di Discussione N.6, ISTAT,Roma.

[35] RUSSO A.(I986),Su un metodo di stima dell'effetto ponderazione nei campioni a due stadi con stratificazione delle unità primarie,Quaderni di Discussione N.I,ISTAT,Roma.

[36] RUSSO A. (I986),Una metodologia per la stima degli effetti stratificazione,clustering,ponderazione e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con selezione delle unità primarie con reimmissione e probabilità variabile,Quaderni di Discussione N.2 ,ISTAT, Roma.

[37] RUSSO A. e ZONNO G. (I986),Indagine sulle opinioni e gli atteggiamenti degli italiani sulle tendenze demografiche: Piano della rilevazione campionaria,riporto dei dati all' universo,calcolo e presentazione errori campionari,Working Paper N.I,Istituto di Ricerche sulla Popolazione,C.N.R.,Roma.

[38] RUSSO A.,FALORSI P.D. e COCCIA G.(I986),Indagine speciale sulle letture in Italia nel I984:Disegno di campionamento calcolo e presentazione degli errori campionari,ISTAT,Roma.

[39] SUDMAN S. e BRADBURN N.H.(I974),Response effects in surveys, Aldine Publ.Co.Chicago.

[40] THEIL H.(I977),Principi di econometria,UTET,Torino.

[41] U.S.DEPARTMENT OF COMMERCE (I974),Standard for Discussion and Presentation of Errors in Data,Technical Paper,N.32, Washington.

[42] U.S.DEPARTMENT OF COMMERCE (I984),The Current Population Survey:A Report on Methodology,Technical Paper,N.7,Washington.

[43] U.S.DEPARTMENT OF COMMERCE,(I968),Evaluation and Research Program of the U.S.Censuses of Population and Housing,I960: Effects of Interviewers and Crew Leaders,Series ER 60,N.7, Washington.

[44] VERMA V.,SCOTT C. e O'MUIRCHEARTAIGH (I980),Sample designs and Sampling errors for the World Fertility Survey,Journal Royal Statistical Society,A,Part.4.

[45] VERMA V. (I982),The estimation and presentation of sampling errors,Technical Bulletins,World Fertility Survey,New York.

[46] WOLTER K.M.,(I985),Introduction to variance Estimation,Springer-Verlag,New York.

[47] WOODRUFF R.S.,(I97I),Simple Method for Approximating the variance of a Complicated Estimate,Journal American Statistical Association,Vol.66.

[48] YAMANE T.,(I967),Elementary Sampling Theory,Prentice- Hall Inc. Englewood Cliffs,N.J.

[49] ZANNELLA F. (I982),Indagine sulle condizioni di salute della popolazione italiana e sul ricorso ai servizi sanitari nel I980:Calcolo degli errori di campionamento,Supplemento al Bollettino di Statistica,N.I2,ISTAT,Roma.

[50] ZANNELLA F., (I982), Piano della rilevazione campionaria ed errori di campionamento, in Indagine sulla fecondità in Italia, anno I979, Università di Padova, Firenze, Roma.

[51] ZARKOVICH S.S., (I967), La qualité des données statistiques, F.A.O.

Numeri pubblicati nella collana "Quaderni di discussione":

84.01 REY G.M.
Le statistiche ufficiali e l'attività della Pubblica Amministrazione
Giugno 1984

85.01 CRESCENZI F.
Nota su alcune metodologie per la classificazione di unità territoriali
Febbraio 1985

85.02 CORTESE A.
Alcune considerazioni sulle prospettive del censimento della popolazione
Marzo 1985

85.03. NATURANI G.
Stima delle ore di lavoro effettivamente prestate dai lavoratori occupati negli anni 1960-1983
Aprile 1985

85.04 NAPOLITANO P.
Esposizione di alcune tecniche per la investigazione dei dati
Maggio 1985

85.05 RUSSO A.
Su un metodo di stima degli effetti stratificazione e clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio
Settembre 1985

85.06 RUSSO A. e FALORSI P.
Rilevazioni campionarie delle forze di lavoro. Metodologia del campionamento calcolo e presentazione errori campionari
Novembre 1985

85.07 PAGNANELLI F.
Natimortalità, mortalità perinatale, mortalità infantile nel Comune di Napoli negli anni dal 1976 al 1980
Dicembre 1985

85.08 STEFANUTTI DE SIMONE L.
Le componenti stagionali delle variazioni dei prezzi al consumo dei prodotti alimentari
Dicembre 1985

86.01 RUSSO A.
Su un metodo di stima dell'effetto ponderazione nei campioni a due stadi con stratificazione delle unità primarie
Gennaio 1986

86.2 RUSSO A.
Una metodologia per la stima degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo nel disegno di campionamento nei campioni a due stadi con selezione delle unità primarie con reimmissione e probabilità variabile.
Luglio 1986

87.01 DE NICOLA I. , CECCARELLI M., CALZARONI M.

Nota sulle statistiche nel settore della edilizia e delle opere pubbliche

Gennaio 1987

87.02 MILITELLO A.

Un confronto tra redditi dichiarati al Fisco e redditi stimati dalla Contabilità Nazionale per gli anni 1981 e 1982

Febbraio 1987

87.03 MAROZZA F.

Centenario dell'International Statistical Institute (ISI): adozione delle tecniche informatiche per la statistica

Febbraio 1987

87.04 RUSSO A.

Sulla presentazione egli errori di campionamento mediante modelli.

Il metodo dei modelli regressivi.

Marzo 1987