

2- 2005



SISTEMA STATISTICO NAZIONALE  
ISTITUTO NAZIONALE DI STATISTICA

TECNICHE E STRUMENTI

# GENESEES V. 3.0

## Funzione Riponderazione

*Manuale utente  
e aspetti metodologici*



 Istat



SISTEMA STATISTICO NAZIONALE  
ISTITUTO NAZIONALE DI STATISTICA

# **GENESEES V. 3.0**

Funzione Riponderazione

*Manuale utente  
e aspetti metodologici*

*A cura di:* Daniela Pagliuca  
e-mail: pagliuca@istat.it

*Sezione I:* Cap. 1, Cap. 2, Cap. 3 Daniela Pagliuca; Cap. 4 Stefano Falorsi; Cap. 5 Daniela Pagliuca e Patrizia Giaquinto; Cap. 6 Paolo Righi; Cap. 7 Daniela Pagliuca.

*Sezione II:* Cap. 1 - paragrafo 1.1 è stato redatto da Daniela Pagliuca; paragrafi 1.2, 1.3, 1.4 e 1.5 sono stati redatti da Paolo Righi. Cap. 2 - paragrafi 2.1 e 2.2 Daniela Pagliuca, paragrafo 2.3 Stefano Falorsi

*Appendici:* A.1 e A.2 Stefano Falorsi; A.3 Paolo Righi

## **GENESEES V. 3.0**

Funzione Riponderazione

*Manuale utente e aspetti metodologici*

Istituto nazionale di statistica  
Via Cesare Balbo, 16 - Roma

*Coordinamento editoriale:*  
Piero Crivelli  
Servizio Produzione editoriale  
Via Tuscolana, 1788 - Roma

*Progetto grafico e videoimpaginazione:*  
Antonio Maggiorani

*Stampa digitale:*  
Istat - Produzione libraria e centro stampa

Febbraio 2005 – copie 250

Si autorizza la riproduzione ai fini  
non commerciali e con citazione della fonte

## **GENESEES V 3.0**

(GENERalised software for Sampling Estimates and Errors in Surveys)

*Software generalizzato per il calcolo dei pesi, delle stime e degli errori campionari*

Genesees V. 3.0 è un software generalizzato nato da diverse procedure SAS, sviluppate da Piero Demetrio Falorsi e Stefano Falorsi, per il calcolo dei pesi e delle stime mediante stimatori di regressione generalizzata, per il calcolo degli errori campionari, per la loro presentazione sintetica mediante modelli regressivi. Tali procedure, dal punto di vista dell'architettura e degli algoritmi utilizzati, costituiscono la base delle funzioni di "Riponderazione" e di "Stima ed Errori campionari" attualmente disponibili anche in Genesees V. 3.0; rispetto alla versione 2.0 il software Genesees V.3.0 comprende una funzione aggiuntiva, la funzione Analisi dei Modelli, che agevola l'utente nella rappresentazione sintetica degli errori campionari, permettendo la visualizzazione grafica dei dati per tenere in considerazione e eventualmente eliminare i valori estremi. Genesees V.3.0 è stato realizzato all'interno di un progetto di sviluppo dell'unità MTS/F "Software generalizzati per la produzione statistica" dell'Istat, responsabile Daniela Pagliuca, in collaborazione con l'unità PSM / A "Strategia campionaria e tecnica di rilevazione", responsabile Stefano Falorsi. Il progetto ha avuto come obiettivo quello di ottimizzare le procedure SAS, implementando i controlli necessari per l'esecuzione e sviluppando una interfaccia user-friendly per consentire agli utenti un'interazione di tipo avanzato, e di implementare ex-novo la funzione Analisi Modelli. Stefano Falorsi è il responsabile delle metodologie statistiche implementate nel software. Si ringraziano Piero Falorsi e Giulio Barcaroli per i commenti ed i suggerimenti.



# Indice

<b>Presentazione</b>	9
 <b>SEZIONE I:</b> <b>Il software Genesees V. 3.0 e la funzione di Riponderazione</b>	
<b>1. Introduzione: cosa contiene questo manuale e come utilizzarlo</b>	17
1.1 Cosa contiene il manuale	17
1.2 Come utilizzare il manuale: alcune indicazioni sui capitoli	18
<b>2. L'installazione e l'avvio del software</b>	23
2.1 I requisiti hardware e software e modalità di installazione	23
2.2 La procedura di avvio e la password di esecuzione	25
2.3 Assistenza al software	26
<b>3. Il software Genesees V. 3.0: un insieme di funzioni</b>	31
3.1 La struttura del software Genesees V. 3.0	31
3.2 Le funzioni del software Genesees V. 3.0	34
<b>4. La funzione di Riponderazione: cenni metodologici</b>	37
<b>5. L'utilizzo della funzione di Riponderazione del software Genesees V. 3.0</b>	47
5.1 La schermata principale	47
5.2 Il calcolo dei pesi finali	49
5.2.1 <i>Le variabili e i parametri di input</i>	53
5.2.2 <i>La selezione delle variabili di input tramite le maschere di selezione</i>	54

5.2.3 <i>La selezione delle variabili di input tramite i parametri attivati dal software</i>	59
5.3 La funzione "Crea stampe"	61
<b>6. La descrizione delle stampe prodotte dalla funzione di Riponderazione del software Genesees V. 3.0</b>	<b>71</b>
6.1 Tavola 1 - Statistiche sulle stime e pesi finali per popolazione Pianif. Stimatore	71
6.2 Tavola 2 - Statistiche sui correttori per popolazione Pianif. Stimatore	73
6.3 Tavola 3 - Statistiche sulle stime e sui pesi diretti per popolazione Pianif. Stimatore	75
6.4 Tavola 4 - Parametri prefissati per la procedura iterativa di stima	77
6.5 Tavola 5 - Totali noti, stime dirette, stime finali e differenze	78
6.6 Tabulati di controllo	80
<b>7. I file di output della funzione di Riponderazione di Genesees</b>	<b>83</b>

## SEZIONE II:

### Approfondimenti sulla costruzione dell'input e sui data-set di output della funzione di Riponderazione

<b>1. La costruzione del data-set di input della funzione di Riponderazione</b>	<b>87</b>
1.1 Le variabili ed i parametri di input	87
1.1.1 <i>Le variabili di input</i>	88
1.1.2 <i>I parametri di input e la convergenza della procedura</i>	92
1.1.3 <i>I controlli sui data-set di input</i>	95
1.2 Definizione delle variabili di input per un dato stimatore	96
1.2.1 <i>Gruppo di riferimento del modello</i>	98
1.2.2 <i>Stimatore definito sui totali noti di una sola variabile ausiliaria</i>	101
1.2.3 <i>Stimatore definito su diversi sistemi di totali noti</i>	112
1.2.4 <i>Scelta dello schema di costruzione del data-set TOTINP e INP</i>	114
1.3 Definizione delle variabili ausiliarie e della variabile "peso distanza" per calcolare i coefficienti finali di alcuni importanti stimatori campionari	116

1.4 Scelta della funzione di distanza per definire i coefficienti finali di alcuni stimatori campionari	126
1.5 Utilizzo del software per il trattamento delle mancate risposte totali	127
<b>2. L'output del software</b>	133
2.1 Il data-set dei parametri di input	133
2.2 Gli errori rilevati sul data-set di input	134
2.2.1 I controlli sul data-set dei totali noti e la scrittura del data-set NOTI_MISS	134
2.2.2 I controlli sul data-set dei campionari e la scrittura del data-set MISSING e del data-set CODICI DOPPI	135
2.2.3 Il controllo della corrispondenza tra i valori dei due data-set di input e la scrittura dei data-set CSENZAT e VUOTI	135
2.3 I data-set contenenti i pesi finali	136
<b>APPENDICI</b>	
<b>A.1 Stimatore di ponderazione vincolata</b>	141
A.1.1 Premessa	141
A.1.2 Caratteristiche generali degli stimatori di ponderazione vincolata	143
A.1.3 Scelta della funzione di distanza	149
<b>A.2 Stimatore di regressione generalizzata</b>	159
A.2.1 Premessa	159
A.2.2 Modello a livello di unità elementari	161
A.2.2.1 Simbologia e prima formulazione dello stimatore di regressione generalizzata	161
A.2.2.2 Espressioni alternative dello stimatore di regressione generalizzata	167
A.2.2.3 Alcune considerazioni sul ruolo del modello	169
A.2.3 Livello del modello	170
A.2.3.1 Introduzione al problema	170
A.2.3.2 Campionamenti a grappoli	172
A.2.3.3 Disegni di campionamento a due o più stadi	177
A.2.4 Gruppo di riferimento del modello	181
A.2.4.1 Modello a livello di unità elementare	181
A.2.4.2 Modello a livello di grappolo	187
A.2.5 Tipo di modello	189



<b>A.3 La costruzione dei data-set di input per definire i gruppi di riferimento</b>	197
A.3.1 Costruzione dei gruppi di riferimento: caso I	197
A.3.2 Costruzione dei gruppi di riferimento: caso II	211
<b>Bibliografia</b>	215

# Presentazione

Le indagini campionarie condotte da un ente produttore di statistiche ufficiali, quale è l'Istituto Nazionale di Statistica, sono finalizzate all'ottenimento di stime di parametri (totali, medie, frequenze, rapporti) nella popolazione di interesse.

Il calcolo di tali stime è reso possibile dall'adozione di particolari funzioni dei valori assunti dalle unità appartenenti ai campioni rilevati: tali funzioni sono dette “stimatori”.

Dato un campione probabilistico, ad ogni unità del quale è associata una precisa e calcolabile “probabilità di inclusione”, dipendente dal disegno campionario adottato, gli stimatori più semplici da utilizzare sono quelli noti come stimatori di espansione, o di Horvitz-Thompson

Ogni stimatore fa uso del peso associato alle unità campionarie: il peso (o coefficiente di espansione o di riporto all'universo) indica quante altre unità della popolazione (non rientranti nel campione) vengono rappresentate dall'unità campionaria cui esso si riferisce. Il calcolo delle stime avviene semplicemente “contando” ogni unità campionaria tante volte per quante unità nella popolazione essa rappresenta.

Nel caso dello stimatore di Horvitz-Thompson, il peso altro non è che l'inverso della probabilità di inclusione dell'unità. Poiché questo peso è definito dal disegno del campione, una volta raccolti i dati non sono necessarie altre operazioni ed è possibile procedere direttamente al calcolo delle stime. L'unica eccezione si ha quando si riscontra una presenza non trascurabile di mancate risposte totali: per farvi fronte, la singola

pionaria effettivamente rispondente deve rappresentare anche parte delle unità rientranti nel campione teorico, ma non in quello effettivo. In tal caso, il peso da utilizzare nel processo di calcolo delle stime sarà sempre l'inverso della probabilità di inclusione, moltiplicato per un opportuno fattore di correzione necessario per tener conto del fenomeno della mancata risposta.

Lo stimatore di Horvitz-Thompson gode della proprietà della correttezza: la media delle stime ottenute mediante la sua applicazione a tutti i possibili campioni estraibili dalla popolazione coincide col valore vero del parametro oggetto di inferenza (in tal caso la distorsione è nulla). Tale stimatore può però non essere il più efficiente, quello cioè mediante il quale si consegue la minima varianza delle stime nell'universo dei campioni: altri stimatori possono garantire tale risultato.

Poiché l'accuratezza delle stime è valutata mediante l'errore quadratico medio (Mean Square Error, MSE), che è dato dalla somma di componenti quali distorsione e variabilità, allo stimatore di Horvitz-Thompson possono essere preferiti altri stimatori che, anche se non corretti, o corretti solo asintoticamente (cioè al di sopra di una certa dimensione campionaria), garantiscono un migliore MSE in quanto producono un guadagno notevole in termini di efficienza (cioè di riduzione della variabilità). Tali stimatori fanno normalmente uso di informazione ausiliaria, di informazione cioè disponibile per l'intera popolazione (a livello di singole unità o di totali) correlata a quella direttamente necessaria per la produzione delle stime (e disponibile mediante le osservazioni campionarie).

Gli stimatori di regressione generalizzata costituiscono un esempio classico di stimatori che fanno uso di informazione ausiliaria, attraverso la definizione di un modello che lega la variabile oggetto di stima (variabile dipendente) a una o più variabili ausiliarie (indipendenti) i cui valori sono noti nel complesso della popolazione, non necessariamente per ogni singola unità, ma anche solo a livello di totali.

L'uso dello stimatore di regressione generalizzata comporta un processo di riponderazione delle osservazioni campionarie. Il peso iniziale, inverso della probabilità di inclusione, viene moltiplicato per un fattore di correzione da determinare mediante procedura iterativa, tale da assicurare da

una parte la minimizzazione della distanza euclidea tra peso iniziale e peso finale, e dall'altra la coincidenza tra i totali noti nella popolazione ed i totali stessi ottenuti mediante stima calcolata dai valori campionari. Questo risultato, laterale rispetto a quello di un migliore MSE, è comunque estremamente importante per un Istituto di statistica ai fini della coerenza delle stime prodotte da indagini diverse.

Si dimostra che lo stimatore di regressione generalizzata è asintoticamente corretto: in altri termini, le stime ottenibili attraverso di esso non sono soggette a distorsione per campioni la cui dimensione sia sufficientemente elevata (condizione praticamente sempre rispettata dalle indagini campionarie dell'Istat). Per quanto riguarda l'efficienza, essa dipende dalla bontà del modello che lega la variabile oggetto di inferenza con le variabili ausiliarie: quanto maggiore è la variabilità spiegata dal modello adottato (o anche: quanto minori risultano essere i residui, cioè le differenze tra i valori osservati e quelli predetti), tanto maggiore è l'efficienza e tanto minore la variabilità delle stime (e conseguentemente l'MSE).

Lo stimatore di regressione generalizzata è un caso particolare della classe più generale degli stimatori di calibrazione. Ogni stimatore di calibrazione è individuato dalla particolare funzione di distanza tra peso iniziale e peso finale utilizzata nel processo di riponderazione: oltre alla distanza euclidea (utilizzata come già detto per lo stimatore di regressione), sono definite altre distanze, quali la logaritmica, la logaritmica troncata, la lineare troncata, quella di minima entropia ecc. che definiscono altrettanti stimatori di calibrazione, la cui scelta è legata a particolari esigenze di indagine.

Il software per la riponderazione delle osservazioni campionarie finalizzato all'utilizzo di stimatori di calibrazione nasce dallo studio e dalle attività di alcuni ricercatori del Servizio Studi Metodologici dell'Istat negli anni '80 (in primo luogo Piero D. Falorsi e Stefano Falorsi, che oggi dirigono il Servizio *Progettazione e Supporto Metodologico nei processi di produzione statistica*). Il Servizio Studi, anche in quegli anni, garantiva la copertura delle fasi peculiari delle indagini campionarie: da un lato, la progettazione del campione (definizione della dimensione, degli strati, allocazione delle unità, scelta delle modalità di selezione), dall'altro l'elaborazione delle stime e la valutazione del grado di affidabilità di queste. In particolare,

uno degli obiettivi era di disporre di strumenti che permettessero di coprire integralmente questa seconda fase. Per primo, fu sviluppato un prototipo software per calcolare i pesi campionari finali tenendo conto di totali noti della popolazione oggetto di studio e garantendo la coincidenza tra questi e le corrispondenti stime campionarie. Immediatamente dopo venne implementato un secondo prototipo per calcolare le stime e gli errori campionari. I due prototipi sono stati sviluppati da Piero Falorsi e Stefano Falorsi (Falorsi P. e Falorsi S., 1995; Falorsi P. e Falorsi S., 1997). La disponibilità di tali strumenti permise di trattare in modo efficace ed omogeneo le fasi di elaborazione dei dati campionari, relativamente alle più importanti indagini condotte dall'Istituto. Con un limite, però: le caratteristiche dei due sistemi, dal punto di vista di facilità di utilizzo, non erano tali da permetterne un uso agevole ad utenti che non fossero quelli già esperti del Servizio Studi. Data la scarsità di risorse in tale Servizio, il trattamento di più indagini in parallelo è stato spesso difficoltoso. Nell'ottica poi di estendere l'applicazione delle tecniche di calibrazione e di valutazione della varianza campionaria anche alla più vasta utenza potenziale del SISTAN, si comprende come questo limite diventasse difficilmente accettabile.

Per tale motivo, ed anche al fine di ottimizzare l'efficienza elaborativa degli algoritmi interni, si decise di sviluppare software di uso generale, partendo dai suddetti prototipi. Si scelse di procedere con lo sviluppo interno - anziché utilizzare procedure statistiche disponibili presso altri enti statistici o prodotti di mercato - per due motivi: da un lato, assicurare al software le stesse caratteristiche metodologiche già implementate nei prototipi, di cui era nota la capacità di soddisfare le esigenze della quasi totalità delle indagini ISTAT, caratterizzate da un'alta complessità delle strategie campionarie adottate. Dall'altro, garantirsi la possibilità di poter intervenire in qualsiasi momento ed in piena autonomia al fine di arricchire i sistemi con le tecniche innovative che la ricerca costantemente produce in questo settore.

Il software Genesees è stato sviluppato in diverse fasi progettuali dall'unità che si occupa di software generalizzato per la produzione statistica (attualmente, collocata nel Servizio *Metodologie, Tecnologie e Software* con la denominazione MTS/F - "Software generalizzati per la produzione stati-

stica”), la cui responsabilità è stata affidata a Daniela Pagliuca e che ha previsto l’inserimento nel progetto di esperti informatici quali Roberto Di Giuseppe e Marco Landriscina. Genesees V. 1.0 è nato come un software generalizzato per il calcolo delle stime e degli errori campionari; la versione 2.0 ha integrato in tale software la funzione di Riponderazione, per il calcolo dei pesi finali. Infine l’attuale versione del software – Genesees V. 3.0 – garantisce, oltre alle due funzioni citate sopra, anche quella per la stima e l’analisi dei modelli per la presentazione degli errori campionari, funzione implementata ex-novo nell’ambito di un progetto diretto dall’unità MTS/F, per agevolare l’utente nella rappresentazione sintetica degli errori campionari, permettendo la visualizzazione grafica dei dati per individuare, ed eventualmente eliminare, i valori *estremi*.

La funzione cui questo manuale si riferisce è quella di **Riponderazione**, contenuta in Genesees V. 3.0 (GENeralised Sampling Estimates and Errors in Surveys).

**Giulio Barcaroli**

*Responsabile del Servizio  
Metodologie, tecnologie e  
software per la produzione statistica*

**Piero Demetrio Falorsi**

*Responsabile Servizio  
Progettazione  
e Supporto Metodologico*



# SEZIONE I

**Il software Genesees V. 3.0  
la funzione di Riponderazione**





# 1. Introduzione: cosa contiene questo manuale e come utilizzarlo

Il presente manuale guida gli utenti che devono fare uso della **FUNZIONE DI RIPONDERAZIONE** del software **Genesees V. 3.0**.

In particolare:

- Aiuta l'utente ad installare il software Genesees V. 3.0, evidenziando i requisiti hardware e software richiesti;
- Descrive la struttura del software Genesees V. 3.0 nel suo complesso, come insieme di funzioni;
- Descrive la metodologia che è alla base della funzione di Riponderazione del software Genesees V. 3.0;
- Presenta la funzione di Riponderazione, descrivendo le maschere che possono essere richiamate per il calcolo dei pesi campionari;
- Descrive come costruire l'input appropriato per il calcolo dei pesi campionari e analizza i dati di output;
- Illustra le stampe ottenibili tramite la funzione di Riponderazione.

## 1.1 Cosa contiene il manuale

Il manuale è diviso in due sezioni e comprende inoltre una appendice metodologica.

La **Sezione I** costituisce il manuale vero e proprio per **l'utilizzo della funzione di Riponderazione**: in essa è descritta la base metodologica, vengono illustrate le schermate presentate dal software e le stampe.

Gli approfondimenti relativi ai dati di input e di output vengono demandati alla sezione successiva.

I primi tre capitoli sono introduttivi al software Genesee V. 3.0: il presente *capitolo 1* illustra il contenuto del manuale e la modalità di utilizzo; il *capitolo 2* aiuta l'utente ad installare ed avviare il software e il *capitolo 3* si riferisce al software nel suo complesso.

Dopo i primi tre capitoli introduttivi al software, il manuale descrive in dettaglio la funzione di Riponderazione.

La **Sezione II** approfondisce i dati di input e output della funzione di Riponderazione, illustrando dettagliatamente come costruire il data-set di input e descrivendo i data-set di output. E' da osservare che per utilizzare la funzione di Riponderazione è richiesta la costruzione di un data-set di input e tale operazione deve effettuarsi seguendo criteri ben definiti. La configurazione del data-set di input è perciò trattata come approfondimento nella Sezione II, in quanto è rivolta a chi, avendo una adeguata preparazione metodologica, è in grado di comprendere le scelte metodologiche sottostanti la strategia campionaria. Anche i data-set di output sono approfonditi in questa seconda sezione.

L'**APPENDICE** infine approfondisce gli aspetti metodologici alla base della funzione di Riponderazione.

## **1.2 Come utilizzare il manuale: alcune indicazioni sui capitoli**

Per agevolare l'utilizzo del software vengono di seguito riportate alcune indicazioni utili per l'utente, descrivendo quanto riportato nei capitoli del manuale.

La Sezione I è formata dai Capitoli 1, 2, 3, 4, 5, 6, 7.

La Sezione II è formata dai Capitoli 1, 2.

All'interno del manuale, il richiamo ad altri capitoli o paragrafi (quale ad esempio: *cfr. paragrafo 3.2*), ove non venga specificata la sezione di rimando, va inteso riferito alla stessa sezione.

## SEZIONE I

Nella **Sezione I** è riportata la modalità di utilizzo della funzione del software Genesees V. 3.0 per il calcolo dei pesi finali.

Nel dettaglio l'utente può leggere come:

- a) Installare il software
- b) Utilizzare le schermate presentate dal software
- c) Selezionare le stampe desiderate e leggere i dati di output

## Capitolo 1

Il presente *capitolo 1* è introduttivo e illustra il contenuto del manuale e il suo utilizzo.

## Capitolo 2

Il *capitolo 2* descrive la procedura di installazione ed avvio del software Genesees V. 3.0.

Per **installare il software** l'utente riceve un CD-ROM contenente un programma di installazione, le cui informazioni essenziali sono riportate nel *capitolo 2*.

Tali informazioni sono anche disponibili (*aggiornamento 2005*):

- via internet (per utenti esterni all'istat):  
<http://www.istat.it/Metodologi/index.htm> (selezionare "Metodi e Software per indagini statistiche").
- via intranet (per utenti istat):  
<http://intranet/> (selezionare: "Prodotti e Applicazioni. Software Generalizzati") e da qui selezionare "MTS-F: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)".

## Capitolo 3

Dopo l'installazione è utile leggere il *capitolo 3*, introduttivo a Genesees V. 3.0. Il *capitolo 3* infatti presenta il software Genesees V. 3.0 nel suo complesso, come **insieme di funzioni**.

## Capitolo 4

Il *capitolo 4* introduce i **cenni metodologici** alla base della funzione di Riponderazione del software Genesees V. 3.0.

## Capitolo 5

Il *capitolo 5* descrive in dettaglio **come usare le schermate del software Genesees V. 3.0 per il calcolo dei pesi finali** e presenta sommariamente le stampe, approfondite nel successivo *capitolo 6*.

I *paragrafi 5.1 e 5.2* supportano l'utente descrivendo come utilizzare le maschere del software; il *paragrafo 5.3* illustra come produrre le stampe.

In dettaglio:

- Il paragrafo 5.1 è introduttivo e indica come avviare il software, riprendendo quanto già descritto nel capitolo 3 (in riferimento a Genesees V. 3.0, visto nella sua globalità come insieme di funzioni).
- Il paragrafo 5.2 entra nel merito della descrizione dell'uso della funzione di Riponderazione: il paragrafo 5.2.1 introduce le variabili e i parametri di input per la funzione di Riponderazione; nei paragrafi 5.2.2 e 5.2.3 è descritto come selezionare tali variabili di input.; il paragrafo 5.2.4 illustra come eseguire l'elaborazione vera e propria per ottenere i pesi finali.
- L'utente può selezionare le informazioni che desidera ottenere in stampa, scegliendo tra sei possibili output: nel paragrafo 5.3 viene descritta la selezione delle stampe e sono indicate le informazioni che è possibile ottenere.

## Capitolo 6

Le **informazioni contenute nelle stampe** create dalla funzione di Riponderazione sono approfondite nel *capitolo 6*.

## Capitolo 7

Nel *capitolo 7* viene descritto il tipo di output ottenibile dalla funzione di Riponderazione, in termini di file e *data-set*, soffermandosi in particolare sui **file** di output: infatti il software permette di memorizzare le stampe su file ascii ed excel.

## SEZIONE II

Nella **Sezione II** è possibile leggere gli approfondimenti sull'input e output della funzione del software Genesees V. 3.0 per il calcolo delle stime e degli errori campionari.

Nel dettaglio l'utente può leggere come:

- Costruire l'input
- Capire le informazioni contenute nei data-set di output.

## Capitolo 1

Nel *capitolo 1* è possibile approfondire la **costruzione del data-set di input** e, in particolare, si illustra come costruire le variabili sulla base del campione (tipo di stimatore, disegno etc.).

E' necessario:

### 1) Predisporre l'input

Per costruire l'input, l'utente deve essere a conoscenza delle variabili di input da creare e dei parametri richiesti dal software: le variabili e i parametri di input sono descritte nel *paragrafo 1.1.1 e 1.1.2*.

Nel *paragrafo 1.1.3* sono presentati i vincoli da rispettare nel costruire le variabili di input e i controlli che il software effettua in automatico prima dell'esecuzione del data-set di input per segnalare vincoli non rispettati.

- 2) Definire le variabili di input sulla base del tipo di stimatore utilizzato. Nei *paragrafi 1.2 e 1.3* si illustra come definire alcune variabili di input in relazione al tipo di stimatore utilizzato; in particolare, il *paragrafo 1.2.1*, tratta l'identificazione delle variabili che specificano il gruppo di riferimento del modello.
- 3) Definire le variabili di input sulla base della distanza e in presenza di non rispondenti.

Nel paragrafo 1.4 si pone l'attenzione alla determinazione dei valori di alcuni parametri richiamati nelle maschere di lancio del software, ovvero alla scelta della funzione di distanza. Il paragrafo 1.5 approfondisce l'uso del software per calcolare i coefficienti finali di riporto in presenza di unità campionate non rispondenti.

## Capitolo 2

Nel *capitolo 2* sono illustrati dettagliatamente i principali **data-set di output** del software, creati dalla funzione di Riponderazione.

Tra i vari data-set di output:

- un *data-set* è creato dall'elaborazione per memorizzare parametri di input (*paragrafo 2.1*);
- alcuni *data-set* memorizzano eventuali errori rilevati sull'input (*paragrafo 2.2*);
- altri *data-set* contengono le informazioni sui pesi finali (*paragrafo 2.3*).

Ogni capitolo del manuale - e paragrafo ove necessario - è introdotto da una sintesi che aiuta l'utente ad orientarsi nell'uso del manuale stesso.

**Per chiarimenti sull'utilizzo del manuale e del software si può utilizzare l'indirizzo di posta elettronica [mts-f@istat.it](mailto:mts-f@istat.it).**

## 2. L'installazione e l'avvio del software

***Sintesi:** In questo capitolo vengono riportati i requisiti hardware e software richiesti da Genesees V. 3.0 ed è riportata la procedura d'installazione e quella di avvio del software. Le informazioni riportate si riferiscono alla data di pubblicazione.*

### 2.1 Requisiti hardware e software e modalità di installazione

Genesees è un software sviluppato utilizzando il SAS SYSTEM v.8.1 per Microsoft Windows, ovvero un package di uso generale che incorpora statistiche e procedure di analisi dei dati. Per utilizzare Genesees è necessario che sia installato il sistema SAS versione 8 ed in particolare i moduli: **SAS Language and Macro-facility, SAS IML Language, SAS STAT, SAS GRAPH.**

Lo spazio sul disco fisso necessario per l'installazione è di circa 4 MB ed è consigliabile una memoria di almeno 64 MB. Il tempo d'esecuzione della procedura è legato, ovviamente, alla velocità del processore installato e alla dimensione e complessità dei dati da elaborare.

L'utente riceve un CD-ROM di installazione corredato di un programma per installare il software.

Il software è disponibile anche effettuando il download :

- via internet (per utenti esterni all'istat):  
<http://www.istat.it/Metodologi/index.htm> (selezionare “Metodi e Software per indagini statistiche”).
- via intranet (per utenti istat):  
<http://intranet/> (selezionare: “Prodotti e Applicazioni. Software



Generalizzati” e da qui selezionare “MTS-F: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)”.

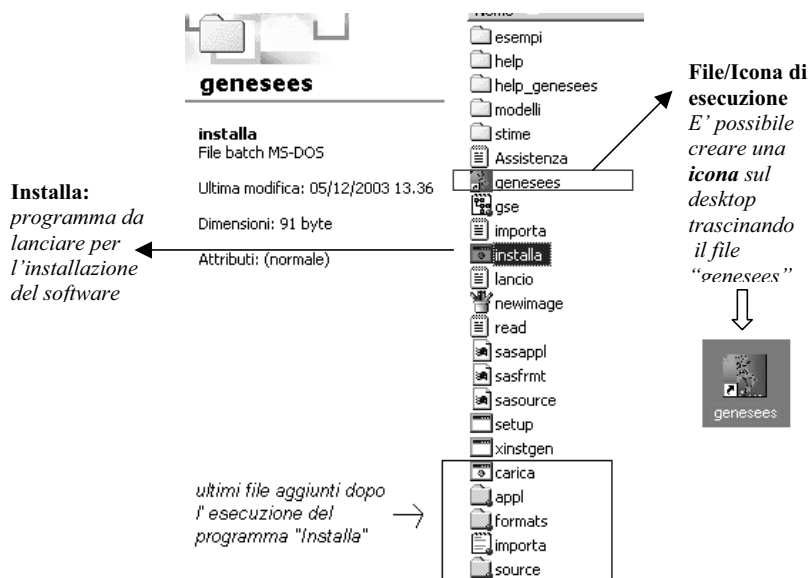
Propedeutica all'installazione del software è - ovviamente - quella del SAS v.8.1

Sia il CD-ROM che il download del software permettono di ottenere un file compresso. Per proseguire con l'installazione è perciò necessario avere a disposizione un programma per espandere il file `genesees3.zip` nella cartella `c:\genesees`.

**Attenzione:** Il file `genesees3.zip` deve espandersi solo nella cartella `c:\genesees`; non è possibile variare il nome della cartella di installazione. Inoltre è necessario installare il software su ogni postazione di lavoro con la procedura di seguito descritta e non è consentito copiare i file, senza effettuare la procedura d'installazione.

La procedura di installazione richiede la sola esecuzione del file **Installa.bat**, che crea nuovi file necessari all'esecuzione dei programmi. Al termine dell'esecuzione la cartella contiene i file mostrati in *figura 2.1*.

**Figura 2.1: Il contenuto della cartella `c:\genesees` d'installazione - successivamente all'esecuzione del programma “`installa.bat`”**



Dopo l'espansione, nella cartella c:\genesees sarà disponibile il file **read.me**, che contiene le istruzioni da eseguire per l'installazione ed il file **Assistenza.txt**, in cui leggere informazioni utili per ricevere assistenza sull'uso del software (*cfr. paragrafo 2.3*).

## 2.2 La procedura di avvio e la password di esecuzione

L'esecuzione del programma installa.bat deve essere effettuata anche nel caso di installazioni successive alla prima.

Una volta installato il programma, la cartella c:\genesees contiene **il file di collegamento “genesees”**, che può essere spostato sul desktop per creare l'icona di lancio (*cfr. figura 2.1*). Il software si avvia perciò cliccando due volte sul file di collegamento "genesees" oppure utilizzando l'icona creata sul desktop.

**Attenzione: nel file collegamento (o nelle proprietà dell'icona) può essere necessario modificare i riferimenti al SAS.**

Infatti, per default il Collegamento che è nelle Proprietà del file o dell'icona, ha la seguente **Destinazione** :

**"C:\Programmi\SAS Institute\Sas\V8\sas.exe" -nologo  
-config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas.**

Se l'utente - ad esempio - ha installato il SAS nel disco D, dovrà cambiare il percorso del file Sas.exe (attenzione: non quello del file lancio.sas o del file gse.cfg, che devono sempre essere riferiti alla cartella c:\genesees).

In dettaglio, il percorso aggiornato deve essere il seguente:

**"D:\Programmi\SAS Institute\Sas\V8\sas.exe"-nologo  
-config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas.**

La proprietà del file di collegamento o della icona sul desktop si varia utilizzando il bottone destro del mouse. Tra le voci che appaiono, selezionare *“Proprietà”* e poi *“Collegamento”*, dove si legge, nel campo *“Destinazione”* il percorso di cui sopra.

Una volta installato, alla prima esecuzione Genesees chiede all'utente di

contattare la struttura che si interessa dello sviluppo e della distribuzione del software generalizzato per ricevere una **password**.

In Istat, l'unità MTS/F si occupa dello sviluppo e distribuzione dei software generalizzati a supporto della produzione statistica nell'ambito del servizio Metodologie e Tecnologie e Software per la Produzione dell'Informazione Statistica (MTF).

Per garantire una tempestiva risposta alle esigenze dell'utenza (sia per ciò che concerne i problemi tecnici che per una veloce e controllata diffusione di password e aggiornamenti), l'unità ha messo a disposizione il seguente indirizzo di posta elettronica : [mts-f@istat.it](mailto:mts-f@istat.it)

Il software dunque mostra una maschera che riporta un codice numerico e richiede la password di registrazione e per riceverla, è necessario contattare l'indirizzo di cui sopra, indicando il codice numerico.

#### **Attenzione:**

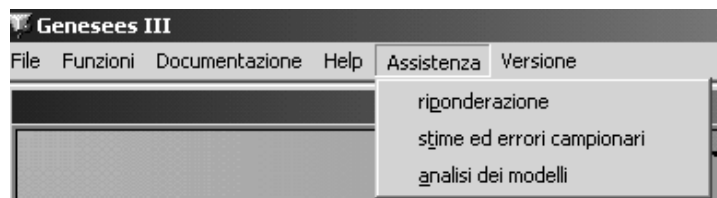
**tale password è a servizio dell'utenza: in tal modo è possibile tenere traccia della lista degli utenti e, di conseguenza, inviare loro eventuali aggiornamenti del software.**

*Dopo la prima esecuzione, le successive installazioni per aggiornamento del software non richiederanno nuove password.*

## **2.3 L'assistenza al software**

Sia nel file Assistenza.txt che è nella cartella c:\genesees che nella voce **Assistenza** della schermata principale (cfr. figura 2.2) vengono riportate alcune informazioni utili all'utente che utilizza il software Genesees.

**Figura 2.2: La voce Assistenza nel Menu di Genesees V. 3.0**



Tali informazioni sono anche evidenziate nelle pagine intranet e internet agli stessi indirizzi riportati nel *paragrafo 2.1* che permettono di effettuare il download del software.

Le informazioni generali e quelle specifiche della funzione di Riponderazione vengono di seguito riportate.

## **INFORMAZIONI SU PROBLEMATICHE RICORRENTI e CONTATTI:**

### **Problemi relativi alla versione dei dataset di input**

Assicurarsi di aver creato tutti i dataset di input in versione SAS V8 (estensione SAS7BDAT oppure SD7). I dataset in versione SAS V6 (SD2) non sono gestiti correttamente dal software.

### **Problemi connessi al funzionamento del software**

#### **Errori durante l'installazione**

1. Ad installazione terminata verificare il log di nome IMPORTA presente nella cartella c:\genesees ed accertarsi che tutti i passi siano terminati con successo. In caso di errori segnalati nel log si consiglia di ripetere tutta la procedura di installazione (non è necessario disinstallare quanto già installato).
2. Se l'installazione è terminata con successo ma non si riesce ad avviare il software, è possibile che nel FILE DI COLLEGAMENTO (o nella icona di collegamento) sia necessario modificare i riferimenti al SAS.

Utilizzando il bottone di destra del mouse sul file di collegamento, è possibile andare in “Proprietà “ e da qui in “Collegamento “, dove si legge la “Destinazione “:

“C:\Programmi\SAS Institute\Sas\V8\sas.exe “ -nologo -config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas “.

Questo percorso deve essere modificato se il SAS è installato su disco o cartelle diverse da quelle di default (ad esempio se in SAS è installato sul disco D).

## FUNZIONE DI RIPONDERAZIONE

### Errori sulla creazione dei dataset di input

Controllare sempre i dataset contenenti le varie tipologie di errore intercettate e segnalate dal software, per avere informazioni sul tipo di errore commesso. In particolare, il software scrive i dataset NOTI\_MISS, CODICI\_DOPPI, CSENZAT, MISSING ove memorizza gli errori rilevati sull'input.

### File di log

Sia nel caso di avvertimento che di errore, uscendo dalla procedura si deve consultare il file `genesees.log` presente nella cartella di output, che contiene appunto il log della elaborazione effettuata.

Se la procedura termina con errore e nel log appaiono messaggi relativi ad un valore esponenziale non ammesso, ciò potrebbe dipendere dall'algoritmo alla base del calcolo dei pesi finali, che non riesce a raggiungere la convergenza.

Per informazioni a tale riguardo contattare gli esperti per problematiche metodologiche (vedi “CONTATTI in Istat “)

Nel caso di procedura terminata con avvertimento, e nel log si leggono informazioni del tipo:

```
%vedi;  
0  
DATA Inptime.Inpass;  
MERGE Inptime.Inpass B;  
BY dominio cod;  
run;
```

WARNING: Multiple lengths were specified for the BY variable dominio by input data sets. This may cause unexpected results.

Tale messaggio non è un errore (attenzione, può divenire fonte di errore solo se l'utente ha creato una variabile “Popolazione utilizzata per lo stimatore “ con una lunghezza maggiore di 15 caratteri, in quanto ciò non è supportato dal software).

## CONTATTI IN ISTAT

### **Per ricevere assistenza sul software GENESEES (aggiornamento 2005)**

**Per ERRORI imputabili al software** (*non relativi alla creazione dei data-set di input*) inviare una e-mail circostanziata, allegando i file di input, il dataset SAVESTIME e il log della elaborazione a:

Roberto Di Giuseppe - Unità Software Generali per la produzione statistica - MTS / F - [digiusep@istat.it](mailto:digiusep@istat.it)

E' preferibile inviare una copia del messaggio anche all'indirizzo:

[mts-f@istat.it](mailto:mts-f@istat.it)

(Indirizzo Operativo dell' Unità Software Generali per la produzione statistica - MTS / F)

**Per problemi connessi con l'INSTALLAZIONE** utilizzare il seguente indirizzo:

[mts-f@istat.it](mailto:mts-f@istat.it) (Indirizzo operativo dell' Unità Software Generalizzati per la Produzione Statistica - MTS / F)

### **Per problematiche metodologiche**

Periodicamente vengono organizzati dei CORSI sugli aspetti metodologici e di utilizzo del software Genesees.

Per gli aspetti metodologici il responsabile è Stefano Falorsi: [stfalors@istat.it](mailto:stfalors@istat.it)

Per avere informazioni che riguardano la CREAZIONE DEI DATA-SET DI INPUT ed in generale per PROBLEMI NON INFORMATICI i contatti consigliati sono:

Paolo Righi : [parighi@istat.it](mailto:parighi@istat.it)

Fabrizio Solari : [solari@istat.it](mailto:solari@istat.it)

### **Contatti in Istat: Informazioni generali**

Nei precedenti punti è possibile identificare i giusti contatti da utilizzare per informazioni o problematiche riguardanti i software di interesse.

Se tali contatti non fossero quelli richiesti, per ricevere le adeguate indicazioni circa gli esperti informatici e metodologi da contattare, così come per informazioni generali sulle attività di sviluppo software generalizzati per la produzione statistica rivolgersi a:

[pagliuca@istat.it](mailto:pagliuca@istat.it) - Responsabile Unità Operativa MTS/F "Software generalizzati per la produzione statistica".



## 3. Il software Genesees V. 3.0: un insieme di funzioni

***Sintesi:** In questo capitolo viene illustrata la struttura del software Genesees V.3.0 e vengono descritte le prime operazioni che l'utente deve attivare per avviare ed operare con il software.*

### 3.1 La struttura del software Genesees V. 3.0

In questo capitolo viene illustrata la struttura del software Genesees V. 3.0, in modo tale che l'utente abbia una immediata visione del prodotto nel suo complesso. Viene descritta anche la schermata iniziale del software, tramite la quale è possibile selezionare le funzioni che lo compongono.

I prossimi capitoli sono invece dedicati alla specifica trattazione della funzione di Riponderazione, oggetto del presente manuale.

Propedeutica all'uso del software è-ovviamente-l'installazione (*cfr. capitolo 2*).

Genesees V. 3.0 viene attivato tramite il file "genesees" che si trova nella cartella c:\genesees d'installazione o tramite l'icona del programma che è stata creata sul desktop:

---

**Figura 3.1: L'icona di avvio**





Con l'avvio della procedura, si apre la schermata principale (*cfr. figura 3.2*), provvista di un menu, in cui compaiono le seguenti opzioni:

- **File:** per uscire dal software o richiamare una precedente elaborazione
- **Funzioni:** per attivare le funzioni principali del software
- **Documentazione:** per accedere alla documentazione on line, ovvero ai manuali di uso delle funzioni di Riponderazione, Stime ed errori campionari e Analisi dei Modelli.
- **Help:** help-on-line sulla schermata di riferimento.
- **Assistenza:** prospetto riassuntivo dei problemi ricorrenti nell'utilizzo del software e contatti in Istat.
- **Versione:** si riferisce all'ultima versione del software.

---

**Figura 3.2 - La schermata principale**

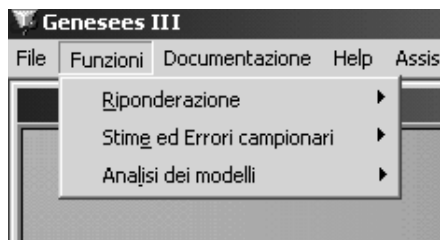


---

L'opzione **Funzioni** fornisce appunto la possibilità di accedere alle tre funzionalità principali implementate nel software (*cfr. figura 3.3*):

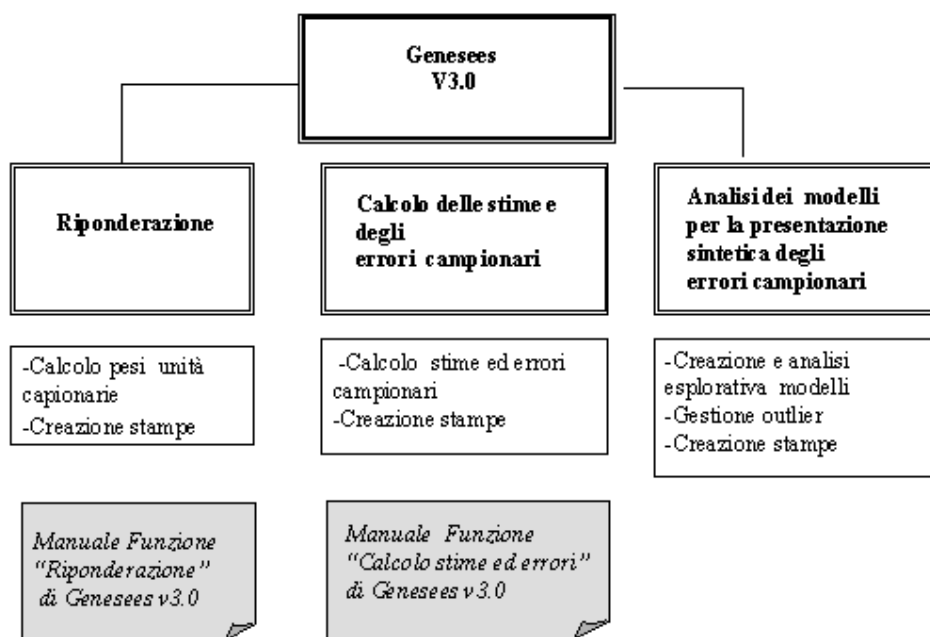
- Riponderazione
- Stime ed Errori campionari
- Analisi dei Modelli

**Figura 3.3 - Le funzioni della schermata principale**



Il software Genesees V. 3.0 è strutturato come mostrato nella successiva *figura 3.4*.

**Figura 3.4 - La struttura del software Genesees V. 3.0**



Come mostrato nella *figura 3.4*, la versione 3.0 comprende tre moduli. La funzione “Analisi dei Modelli” è l'ultimo modulo implementato in aggiunta alla versione 2.0 e tramite questa funzione viene dato ampio spazio alla presentazione grafica degli errori campionari.

Il presente manuale d'uso si riferisce esclusivamente alla funzione di Riponderazione attivata tramite l'opzione "Riponderazione" di *figura 3.3*. Le funzioni di "Stime ed errori Campionari" è descritta in un manuale a se stante (Pagliuca, 2004b).

## **3.2 Le funzioni del software Genesees V. 3.0**

### **La funzione di Riponderazione**

La funzione di Riponderazione è applicabile in tutti i casi in cui esistono informazioni ausiliarie, espresse in termini di totali noti di variabili, definite appunto "ausiliarie", legate alle variabili di interesse.

Essa è finalizzata al calcolo dei pesi finali da attribuire alle unità campionarie, sulla base di totali noti delle variabili ausiliarie e dei valori assunti da queste nel campione estratto.

Il contesto metodologico nel quale la funzione è stata concepita è quello degli stimatori di calibrazione (*calibration estimators*); tale teoria consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di calibrazione (Deville, J. C., Särndal, C. E., 1992, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, vol. 87, pp. 367-382).

### **La funzione di Calcolo Stime ed Errori**

Lo scopo principale delle indagini campionarie è quello di fornire le stime di alcuni parametri descrittivi dell'intera popolazione, o di sottopopolazioni predefinite, dalla quale il campione viene estratto.

La funzione per il calcolo delle stime e degli errori campionari è finalizzata al calcolo delle stime e degli errori di campionamento e produce, per ciascuna sottopopolazione di interesse: le stime oggetto di indagine e i corrispondenti errori di campionamento assoluti, relativi, e gli intervalli di confidenza; le principali statistiche che forniscono informazioni sull'efficienza della strategia di campionamento utilizzata (effetto del disegno ed

effetto dello stimatore); i modelli di regressione per la presentazione sintetica degli errori di campionamento.

Anche tale funzione fa riferimento alla teoria degli stimatori di calibrazione (*calibration estimators*) e della relativa metodologia di calcolo della varianza; la metodologia consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di calibrazione (Deville, J. C., Särndal, C. E., 1992, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, vol. 87, pp. 367-382).

### **La funzione di Analisi dei Modelli**

La funzione di Analisi dei modelli nasce come estensione di quanto già implementato in Genesees V. 2.0 e aiuta l'utente a determinare la migliore rappresentazione sintetica degli errori campionari.

Tale funzione permette infatti di costruire i modelli per la presentazione sintetica degli errori di campionamento, come già era previsto nella versione 2.0 di Genesees, ma permette anche in aggiunta di analizzare la validità di tali modelli, in modo semplice ed interattivo.

La bontà di adattamento dei dati è facilmente migliorabile grazie al supporto di alcune funzionalità grafiche, che agevolano l'utente nel considerare alcuni valori come estremi, e grazie anche alla possibilità di procedere alla determinazione di un nuovo modello, che non tenga in considerazione i valori giudicati *estremi*, senza dover uscire dal software Genesees per modificare i dati di input eliminando i valori estremi.



## 4. La funzione di Riponderazione: Cenni metodologici

La procedura si basa sulla teoria dei *calibration estimators* che nel seguito saranno chiamati stimatori di *ponderazione vincolata*; tale teoria consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di ponderazione vincolata.

Introduciamo la seguente notazione simbolica: sia  $U$  una popolazione finita di  $N$  elementi, che indichiamo con  $U = \{1, \dots, k, \dots, N\}$  e sia  $s$  un campione casuale di  $n$  elementi<sup>1</sup>, che indichiamo con  $s = \{1, \dots, k, \dots, n\}$ , estratto da  $U$  mediante il disegno di campionamento che assegna ad  $s$  la probabilità  $p(s)$  di essere selezionato. Con riferimento alla generica unità  $k \in U$ , indichiamo con  $\pi_k = \sum_{s \ni k} p(s)$  la probabilità di inclusione dell'unità nel campione, con  $y_k$  il valore assunto dalla variabile di interesse  $y$ , con

$$x_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})$$

il valore assunto dal vettore

$$x = (x_1, \dots, x_j, \dots, x_J)$$

di  $J$  variabili ausiliarie.

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

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

sulla base delle seguenti informazioni:

<sup>1</sup>

Per semplicità di esposizione gli elementi dell'universo e del campione sono identificati dalle rispettive etichette.

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

$$X_j = \sum_{k \in U} x_{jk}$$

Sulla base della notazione appena introdotta, uno stimatore del totale  $Y$  appartenente alla classe degli stimatori *di ponderazione vincolata*, può essere espresso mediante la seguente relazione

$$\tilde{Y}_{PV} = \sum_{k \in s} y_k d_k \gamma_{ks} = \sum_{k \in s} y_k w_{ks} \quad (2)$$

in cui  $d_k = \pi_k^{-1}$  per  $k = 1, \dots, n$  indica il *peso diretto* associato alla  $k$ -esima unità estratta,  $w_{ks} = d_k \gamma_{ks}$  denota il *peso finale* associato a tale unità, dove  $\gamma_{ks}$  rappresenta il correttore del peso diretto. Tale espressione è formalmente simile a quella dello stimatore di Horvitz-Thompson (Horvitz D.G. et al., 1952) che, come è noto, è espresso come somma pesata (con i pesi diretti) dei dati campionari. Occorre, tuttavia, far notare che tra i due stimatori esiste una differenza sostanziale in quanto mentre i pesi diretti dipendono unicamente dalle unità estratte nel campione, i pesi finali, come vedremo in seguito, dipendono dai totali noti delle variabili ausiliarie, dai valori assunti dalle variabili ausiliarie nel campione estratto, dalla variabilità della variabile oggetto di indagine. Per mettere in luce tale dipendenza e sottolineare, quindi, la differenza tra lo stimatore di Horvitz-Thompson e quello di ponderazione vincolata si è ritenuto opportuno associare a ciascun peso finale l'indice  $s$ .

L'insieme dei pesi finali  $\{w_{ks}; k = 1, \dots, n\}$  è ottenuto come soluzione del seguente problema di minimo vincolato, in cui

$$\min[E_p \left\{ \sum_s G(d_k, w_{ks}) \right\}] \quad (3)$$

è la funzione obiettivo e

$$\sum_{s \in k} w_{ks} \mathbf{x}_k = \mathbf{X} \quad (4)$$

è il sistema di vincoli, dove con  $G_k(w_{ks}; d_k)$  si è indicata una funzione di distanza tra il peso diretto  $d_k$  ed il peso finale  $w_{ks}$ , ovvero una funzione definita sulla variabile  $w_{ks}$  in cui  $d_k$  rappresenta una costante nota (o *parametro*) della funzione stessa. L'obiettivo è, quindi, quello di individuare un insieme dei pesi finali  $\{w_{ks}; k = 1, \dots, n\}$  che consenta di rispettare il sistema di vincoli (4) e che contemporaneamente modifichi il *meno possibile*, sulla base della funzione di distanza prescelta, l'insieme dei pesi diretti  $\{d_k; k = 1, \dots, n\}$ . Affinché il problema di minimo vincolato, definito dalla (3) e dalla (4), ammetta soluzione e che tale soluzione sia unica, la funzione di distanza  $G_k(w_{ks}; d_k)$  e la sua derivata prima rispetto a  $w_{ks}$ , che indichiamo con

$$g_k(w_{ks}; d_k) = \frac{\delta G_k(w_{ks}; d_k)}{\delta w_{ks}}$$

devono soddisfare alcune condizioni di regolarità che sono indicate nel lavoro di Deville e Särndal (1992). Tali condizioni garantiscono, inoltre, che esista la funzione inversa,  $g_k^{-1}(\cdot)$ , ovvero una funzione per la quale vale

$$w_{ks} = g_k^{-1}(g_k(w_{ks}; d_k)).$$

Al fine di ottenere il vettore  $\mathbf{w} = (w_{1s}, \dots, w_{ks}, \dots, w_{ns})'$  soluzione del problema di minimo vincolato (3) – (4), si utilizza il metodo dei moltiplicatori di Lagrange che consiste nel risolvere il seguente sistema omogeneo

$$\begin{cases} \frac{\delta L(\boldsymbol{\lambda}, \mathbf{w})}{\delta w_{ks}} = g_k(w_{ks}; d_k) - \mathbf{x}'_k \boldsymbol{\lambda} = 0 & \text{per } k = 1, \dots, n \\ \frac{\delta L(\boldsymbol{\lambda}, \mathbf{w})}{\delta \lambda_j} = \sum_{k \in s} w_{ks} x_{jk} - X_j = 0 & \text{per } j = 1, \dots, J \end{cases} \quad (5)$$

di  $(n + J)$  equazioni nelle  $(n+J)$  incognite  $(\mathbf{w}, \boldsymbol{\lambda})$  in cui  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_j, \dots, \lambda_J)'$  è il vettore dei moltiplicatori di Lagrange e  $L(\boldsymbol{\lambda}, \mathbf{w})$  indica la funzione di Lagrange. Dalle prime  $n$  equazioni del sistema (5) si ottiene mediante semplici passaggi la relazione

$$w_{ks} = d_k F_k(\mathbf{x}'_k \boldsymbol{\lambda}) \quad (6)$$

dove la funzione  $F_k(\mathbf{x}'_k \boldsymbol{\lambda}) = g_k^{-1}(\mathbf{x}'_k \boldsymbol{\lambda})$  rappresenta  $\gamma_{ks}$ , ossia il correttore del peso diretto che è funzione della variabile  $u_k = (\mathbf{x}'_k \boldsymbol{\lambda})$  combinazione lineare del vettore di variabili ausiliarie  $\mathbf{x}_k$  e dei  $J$  valori incogniti del vettore  $\boldsymbol{\lambda}$ .

La (6) non è ancora una relazione operativa in quanto non sono noti i



valori numerici del vettore  $\lambda$ . Al fine di pervenire alla determinazione di  $\lambda$ , introduciamo la (6) nelle ultime J equazioni del sistema (5) ottenendo, mediante semplici passaggi, il seguente sistema di J equazioni nelle J incognite  $\lambda_1, \dots, \lambda_j, \dots, \lambda_J$

$$\phi(\lambda) = X - \tilde{X} \quad (7)$$

in cui

$$\tilde{X} = \sum_{k \in S} d_k x_k \quad (8)$$

e

$$\phi(\lambda) = \sum_{k \in S} d_k x_k (F_k(x'_k \lambda) - 1) \quad (9)$$

Nel caso in cui la  $F_k(x'_k \lambda)$  sia una funzione lineare di  $\lambda$  è possibile ottenere una soluzione esplicita al sistema (7); altrimenti una soluzione numerica al sistema (7) può essere ottenuta in modo iterativo ad esempio mediante il metodo di Newton. Altre soluzioni per specifiche funzioni di distanza sono riportate nel lavoro di Singh e Mohl (1996). Indicando, quindi, con  $\lambda = \lambda^*$  il vettore di J valori soluzione del sistema (7), ottenuto mediante soluzione esplicita oppure mediante un metodo iterativo, e sostituendo i valori così ottenuti nell'espressione (6) è possibile calcolare i pesi finali  $w_{ks}$  (per  $k=1, \dots, n$ ).

Consideriamo adesso alcune fondamentali funzioni di distanza  $G_k(w_{ks}; d_k)$ . In questa sede limitiamo l'esposizione unicamente ai tre metodi principali adottati nelle indagini ISTAT ed utili a risolvere la maggior parte dei problemi di stima che si pongono nelle indagini su larga scala (nell'appendice A.1 e nei lavori di Deville e Särndal (1992) e di Singh e Mohl (1996) sono descritte altre funzioni di distanza).

La prima funzione di distanza è quella euclidea (Cassel C. M. et al., 1976) ed è espressa come

$$G(w_{ks}; d_k) = \frac{(d_k - w_{ks})^2}{q_k d_k} \quad (10)$$

in cui  $1/q_k$  indica un peso non correlato a  $d_k$  assegnato alla k-esima unità.

Nella maggior parte delle applicazioni si utilizza il peso uniforme  $1/q_k = 1$ , ma in alcuni casi può essere conveniente utilizzare pesi  $1/q_k$  variabili. Nel lavoro di Alexander (1987), si dimostra l'utilità dell'adozione dei pesi  $1/q_k$  variabili, per risolvere particolari problemi di sottocopertura.

Lo stimatore di ponderazione vincolata che si ottiene mediante l'utilizzo di tale funzione di distanza coincide con lo stimatore di regressione generalizzata (Särndal C.E. et al., 1992). Il correttore del peso diretto  $d_k$  ottenuto mediante le espressioni (5) – (9) assume la seguente forma:

$$F_k(\mathbf{x}_k' \boldsymbol{\lambda}_*) = 1 + \frac{1}{2} q_k \mathbf{x}_k' \mathbf{a} \quad (11)$$

in cui il vettore  $\mathbf{a}$  ( $J \times 1$ ) è dato dall'espressione

$$\mathbf{a} = \left( \sum_{k \in S} \frac{1}{2} q_k d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} (\mathbf{X} - \tilde{\mathbf{X}}) \quad (12)$$

Si osserva quindi dalla (11) e (12) che il correttore della generica unità si ottiene come una combinazione lineare degli elementi del vettore  $\mathbf{x}_k$  dove i coefficienti sono espressi dagli elementi del vettore  $\mathbf{a}$ ; si tratta quindi di un modello di tipo additivo. Il correttore (11) può assumere valori nell'intervallo  $(-\infty, +\infty)$ .

La seconda funzione di distanza è quella logaritmica (Alexander C. H., 1987) espressa da:

$$G(w_{ks}; d_k) = \frac{w_{ks}}{q_k} \ln \left( \frac{w_{ks}}{d_k} \right) - w_{ks} + d_k \quad (13)$$

in cui “ln” indica il logaritmo naturale in base e. Mediante le espressioni (5) – (9) si ottiene un correttore di tipo esponenziale dato da

$$F_k(\mathbf{x}_k' \boldsymbol{\lambda}) = \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \prod_{j=1}^J \exp(q_k x_{kj} \lambda_j) \quad (14)$$

che può assumere valori nell'intervallo  $(0, +\infty)$ . Il correttore (14) è quindi ottenuto mediante un modello in cui si assume che l'effetto delle variabili ausiliarie sia di tipo moltiplicativo.

La terza funzione di distanza è quella logaritmica troncata (Dewille J. C. et al., 1992) data da

$$G(w_{ks}; d_k) = \frac{d_k}{Aq_k} \left( \frac{w_{ks}}{d_k} - L \right) \ln \frac{\frac{w_{ks}}{d_k} - L}{1 - L} + \frac{d_k}{Aq_k} \left( U - \frac{w_{ks}}{d_k} \right) \ln \frac{U - \frac{w_{ks}}{d_k}}{U - 1} \quad (15)$$

dove  $L$  ed  $U$  sono due costanti (il cui significato sarà esplicitato nelle successive pagine) tali che  $L < 1 < U$  ed

$$A = \frac{L(U - 1)}{(U - L)(U - 1)}. \quad (16)$$

Mediante le espressioni (5) – (9) si ottiene un correttore di tipo logit

$$F_k(\mathbf{x}_k' \boldsymbol{\lambda}) = \frac{L(U - 1) + U(1 - L)\exp(Aq_k \mathbf{x}_k' \boldsymbol{\lambda})}{(U - 1) + (1 - L)\exp(Aq_k \mathbf{x}_k' \boldsymbol{\lambda})}. \quad (17)$$

Anche in questo caso il correttore è di tipo moltiplicativo; tuttavia esso assume valori compresi nell'intervallo  $(L, U)$ .

Esaminiamo, adesso, il problema della scelta della funzione di distanza tra quelle esaminate.

La funzione di distanza euclidea può portare a pesi negativi o nulli in quanto il correttore può variare nell'intervallo  $(-\infty, +\infty)$ ; tali pesi in genere non sono accettabili nella maggior parte delle applicazioni. Tale funzione è quella che richiede minor tempo di elaborazione poiché non necessita di metodi iterativi per la soluzione.

La funzione di distanza logaritmica porta invece a pesi sicuramente positivi che possono essere, tuttavia, estremamente alti ed in genere più elevati di quelli ottenuti con la distanza euclidea.

Il più importante vantaggio della terza funzione di distanza considerata è quello di fornire pesi finali che assumono valori compresi nell'intervallo  $(Ld_k, Ud_k)$  in cui le costanti  $L$  ed  $U$  rappresentano rispettivamente il limite inferiore e superiore dell'intervallo  $(L, U)$  di variazione dei correttori. La scelta dei valori da assegnare alle costanti  $L$  ed  $U$  non è arbitraria poiché esiste un valore massimo teorico di  $L$ ,  $L_{\max}$  (inferiore ad 1), e un valore mini-

mo teorico di  $U_{\min}$  (superiore ad 1), che è possibile assegnare ad  $L$  ed  $U$  affinché il sistema (3) e (4) ammetta soluzione. Una regola approssimata (Verma V., 1995§) per definire  $L_{\max}$  e  $U_{\min}$  è espressa dalle seguenti disuguaglianze

$$L_{\max} < \min\{R_1, \dots, R_j, \dots, R_J\} \quad (18)$$

$$U_{\min} > \max\{R_1, \dots, R_j, \dots, R_J\} \quad (19)$$

dove:

$$R_j = \frac{X_j}{\sum_{k \in S} x_{jk} d_k} \text{ per } (j=1, \dots, J);$$

$L_{\max}$  e  $U_{\min}$  sono, quindi, rispettivamente il *limite inferiore* ed il *limite superiore del più piccolo intervallo teorico* ( $L_{\max}, U_{\min}$ ) di variazione dei correttori intorno ad 1 per il quale il sistema (3) – (4) ammette soluzione. Nella definizione della funzione di distanza (15), pertanto, il valore effettivo,  $L_{\text{eff}}$ , da assegnare alla costante  $L$  deve essere minore o uguale di  $L_{\max}$ , mentre il valore  $U_{\text{eff}}$ , da assegnare ad  $U$  deve essere maggiore o uguale di  $U_{\min}$  in modo da definire un intervallo effettivo ( $L_{\text{eff}}, U_{\text{eff}}$ ) in cui l'intervallo minimo teorico ( $L_{\max}, U_{\min}$ ) è completamente contenuto. Nel software applicativo la scelta di  $L_{\text{eff}}$  e  $U_{\text{eff}}$  è guidata da una maschera interattiva in cui i valori di tali costanti vengono scelti in base alle seguenti relazioni:

$$L_{\text{eff}} = \alpha_L L_{\max} \quad , \quad U_{\text{eff}} = \alpha_U U_{\min}$$

dove  $\alpha_L=0,5$  e  $\alpha_U=1,5$  sono le scelte di *default* per la definizione dei *moltiplicatori*  $\alpha_L$  e  $\alpha_U$ ; in tale maschera l'utente può modificare opportunamente i valori di *default* assegnati ad  $\alpha_L$  e  $\alpha_U$  in modo da restringere o allargare l'intervallo effettivo ( $L_{\text{eff}}, U_{\text{eff}}$ ). E' importante sottolineare che nel caso in cui  $\alpha_L=1$  e  $\alpha_U=1$  l'intervallo effettivo coincide con l'intervallo minimo teorico, inoltre, nella scelta di tali costanti valgono le seguenti restrizioni:

$$\begin{cases} \alpha_L \leq 1 & , & \alpha_U \geq 1 \\ \alpha_L < 1 & \text{ se } & L_{\max} = 1 \\ \alpha_U > 1 & \text{ se } & U_{\min} = 1 \end{cases}$$

che hanno la finalità di impedire la definizione di intervalli effettivi per i quali il sistema (3) e (4) non ammette soluzioni.

Un'altra regola, di tipo empirico, che è spesso applicata per determinare i valori di  $L_{\max}$  ed  $U_{\min}$  consiste nell'attribuire ad  $L$  un valore iniziale prossimo allo zero e ad  $U$  un valore iniziale molto elevato; si procede poi, per tentativi successivi, aumentando  $L$  e diminuendo  $U$  verso l'unità. I valori finali di  $L$  e  $U$  sono quelli più prossimi all'unità per cui il sistema (3) – (4) ammette soluzione. Si osserva, infine, che tanto più i totali noti  $X_j$  (per  $j=1,...,J$ ) differiscono dalle corrispondenti stime dirette, tanto più i rapporti  $R_j$  (per  $j=1,...,J$ ) sono diversi dall'unità, e quindi più forte sarà la correzione da apportare ai pesi diretti sulla base della regola (18) e (19). Le differenze in oggetto possono essere dovute a differenti cause, quali ad esempio, l'effetto delle mancate risposte totali o della sottocopertura della lista di selezione oppure della alta varianza campionaria delle stime dirette costruite su poche unità campionarie. In quest'ultimo caso ha senso ridefinire il sistema (4) dei vincoli, aggregando tra loro quei totali noti le cui corrispondenti stime dirette sono basate su poche unità campionarie.

Come si è detto l'utilizzo della funzione di distanza euclidea porta a definire lo stimatore di regressione generalizzata di cui sono note le proprietà di correttezza asintotica e di consistenza; è possibile inoltre calcolare l'espressione linearizzata della varianza di campionamento. Per tale stimatore si possono definire diversi modelli di regressione lineare che legano le variabili oggetto di indagine alle variabili ausiliarie. Ciò significa, in particolare, caratterizzare il modello regressivo in termini di tre elementi fondamentali: il *gruppo di riferimento del modello*, il *livello del modello*, il *tipo di modello*, di cui di seguito diamo una breve definizione (Estevao V. et al., 1995).

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

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

I *gruppi* rappresentano una partizione della popolazione e per ciascun gruppo si definisce uno specifico modello di regressione.

Risulta chiaro che, nella costruzione del modello lineare si possono utilizzare definizioni alternative dei gruppi di riferimento del modello. E' possibile, infatti, definire i gruppi sia sulla base della partizione della popolazione più *fine* possibile – rispetto alla quale sono noti i totali delle variabili ausiliarie - che sulla base di aggregazioni definite a partire dalla partizione più fine. Un caso particolare si ha quando l'intera popolazione definisce l'unico gruppo di riferimento del modello.

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

Le variabili ausiliarie utilizzate nel modello definiscono il *tipo di modello*. In tal modo è possibile definire tutti i più importanti stimatori; ad esempio, ponendo:

- (a)  $\mathbf{x}_k = 1$  ( $\forall k \in U$ ) corrispondente al modello della media, si ottiene lo stimatore di Horvitz-Thompson;
- (b)  $\mathbf{x}_k = 1/q_k$  ( $\forall k \in U$ ), dove  $x_k$  è una singola variabile a valori positivi, corrispondente al modello del rapporto, si ottiene lo stimatore del rapporto;
- (c)  $\mathbf{x}_k = (1, x_k)'$  con  $1/q_k = 1$  ( $\forall k \in U$ ), corrispondente al modello di regressione semplice con intercetta, si ottiene lo stimatore di regressione.

Analogamente è possibile ottenere diverse forme dello stimatore del rapporto post-stratificato e dello stimatore di tipo ratio-raking.

L'utilizzo della funzione di distanza logaritmica e logaritmica troncata porta a definire stimatori di cui non sono note le proprietà. Non sono, inoltre, ancora chiare le caratteristiche del modello regressivo che lega le variabili oggetto di indagine alle variabili ausiliarie. E' possibile, tuttavia,

ricorrere in questo caso al risultato fondamentale dimostrato nel lavoro di Deville e Särndal (1992), il quale afferma che asintoticamente tutti gli stimatori di *ponderazione vincolata* coincidono con lo stimatore di regressione generalizzato. Per campioni sufficientemente grandi è, quindi, possibile affermare che tutti gli stimatori di *ponderazione vincolata* sono approssimativamente corretti, consistenti ed hanno un'espressione della varianza di campionamento che coincide con quella dello stimatore di regressione generalizzato.

Per la scelta delle variabili ausiliarie e per le proprietà dello stimatore, è possibile, quindi, far riferimento allo stimatore di regressione generalizzata; mentre per il calcolo effettivo dei pesi finali, si può utilizzare la funzione di distanza logaritmica troncata che ha le migliori proprietà in termini di intervallo di variazione dei pesi finali.

## **5. L'utilizzo della funzione di Riponderazione del software Genesees V. 3.0**

***Sintesi:** I paragrafi 5.1 e 5.2 supportano l'utente nell'utilizzo delle maschere del software; il paragrafo 5.3 illustra la produzione delle stampe.*

*In particolare:*

*Il paragrafo 5.1 è introduttivo e spiega come avviare il software, riprendendo quanto già descritto nel capitolo 3 (riferito a Genesees V. 3.0, visto nella sua globalità come insieme di funzioni).*

*Il paragrafo 5.2 entra nel merito della descrizione dell'uso della funzione di Riponderazione: il paragrafo 5.2.1 introduce le variabili e i parametri di input per la funzione di Riponderazione; nei paragrafi 5.2.2 e 5.2.3 è descritta la selezione di tali variabili di input.; il paragrafo 5.2.4 illustra come eseguire l'elaborazione vera e propria per ottenere i pesi finali.*

*L'utente può selezionare le informazioni che desidera ottenere in stampa, scegliendo tra sei possibili output: nel paragrafo 5.3 sono riportate la selezione delle stampe e le informazioni che è possibile ottenere.*

### **5.1 La schermata principale**

Come premesso nel capitolo 3, il software Genesees V. 3.0 viene attivato tramite l'icona del programma posta sul desktop o tramite il file di collegamento "genesees", che si trova nella cartella c:\genesees d'installazione.

L'avvio del programma mostra la schermata principale:



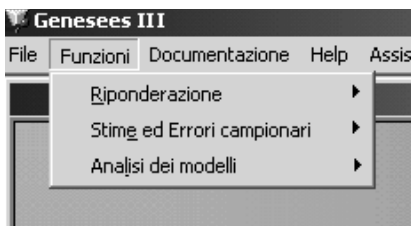
**Figura 5.1 – La schermata principale**



Tramite la voce Funzioni della schermata principale si possono attivare le tre funzioni di:

- Riponderazione
- Stime ed Errori campionari
- Analisi Modelli

**Figura 5.2 – Le funzioni della schermata principale**



In questo manuale viene trattata la **funzione di Riponderazione**, attivata tramite l'opzione omonima.

Come mostrato in *figura 5.3*, essa a sua volta permette di attivare altre due opzioni:

- Calcola pesi finali
- Crea stampe

**Figura 5.3 – La funzione di riponderazione**



L'opzione "Calcola pesi finali" è utilizzata per il calcolo vero e proprio dei pesi finali (*cfr. paragrafo 5.2*); "Crea stampe" crea le stampe relative ad elaborazioni effettuate precedentemente (*cfr. paragrafo 5.3*).

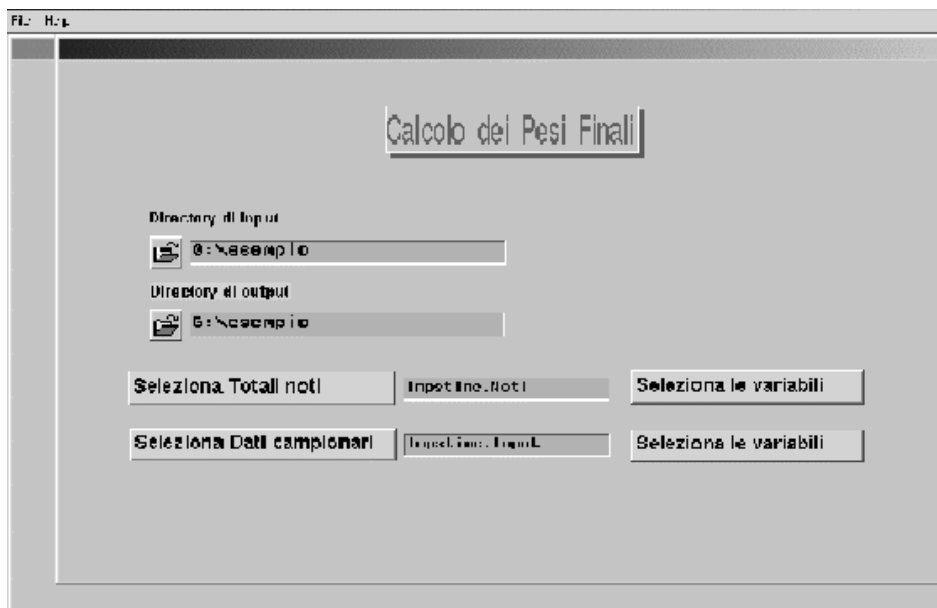
## 5.2 Il Calcolo dei pesi finali

L'opzione "Calcola pesi finali" attiva la maschera  $M_1$  (*cfr. figura 5.4*).

Nella maschera  $M_1$  viene presentato un menu bar, in cui compaiono le seguenti voci:

- **File:** per uscire dalla funzione
- **Help:** per visualizzare l'Help on line relativo alla maschera  $M_1$

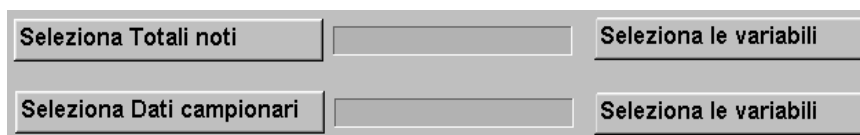
**Figura 5.4 - Maschera di selezione per i dati di input.**



( $M_1$  attivata da  $M_0$ )

Questa maschera consente di effettuare le seguenti scelte:

- la scelta delle cartelle di input e di output: la cartella di input contiene i due *data-set* SAS di input; la cartella di output serve a memorizzare i *data-set* creati dalla procedura e gli eventuali file di stampa;
- la selezione dei due *data-set* contenenti i **totali noti** e i **dati campionari** e delle relative variabili.



La selezione delle variabili di interesse attiva le maschere  $M_2$  e  $M_3$  (cfr. figura 5.5 e 5.6).

**Figura 5.5 – Maschera di selezione delle variabili di input – Variabili del data-set “Totali Noti”**

**TOTALI NOTI**

**Selezionare le Variabili**

☐ Popolazioni pianificate utilizzate per lo stimatore

☐ totali noti delle variabili X1,...,Xn

*(M<sub>2</sub> attivata da M<sub>1</sub>)*

**Figura 5.6 – Maschera di selezione delle variabili di input – Variabili del Data-set “Dati Campionari” e dei parametri di input**

**DATI CAMPIONARI**

**Selezionare le Variabili**

☐ Popolazioni pianificate utilizzate per lo stimatore

☐ Codice identificativo unità campionaria

☐ Variabili ausiliarie

☐ Peso diretto

☐ Peso distanza (opzionale)

Funzioni di distanza

Logaritmica troncata

Logaritmica

Euclidea

Lineare troncata

Hellinger

Minima Entropia

Chi quadrato

0.5

Coeff. moltiplicatore valore minimo di L

1.5

Coeff. moltiplicatore valore massimo di U

Pop. pianif. per stimatore

☐ Riga selezionata

*(M<sub>3</sub> attivata da M<sub>1</sub>)*

In entrambe le maschere  $M_2$  ed  $M_3$  viene presentato un menu bar, in cui compaiono le seguenti voci:

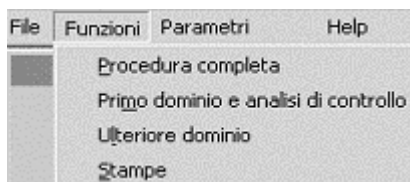
- **File:** per tornare alla maschera principale;
- **Funzioni:** per scegliere l'esecuzione della procedura su tutti i dati di input o su una sua partizione; tale funzione è attiva solo per la maschera  $M_3$ ;
- **Parametri:** per utilizzare i parametri della precedente sessione di lavoro;
- **Help:** per visualizzare l'Help on line relativo alla maschera attiva.

In particolare la voce **Funzioni** attiva le elaborazioni per il calcolo dei pesi finali. E' da osservare che tale voce appare in entrambe le maschere  $M_2$  ed  $M_3$  ma non è consentito eseguire le elaborazioni dalla maschera  $M_2$ . Tramite la maschera  $M_3$  è invece possibile attivare le seguenti quattro voci:

- Procedura completa
- Primo dominio e analisi di controllo
- Ulteriore dominio
- Stampe

---

**Figura 5.7 – Le “Funzioni” attivate dalla maschera  $M_3$**



---

La voce “*Procedura completa*” viene utilizzata per avviare il calcolo di pesi finali – dopo aver selezionato le variabili dei *data-set* di input – allo scopo di creare i *data-set* di output sulla base di tutte le partizioni definite dalla variabile “Popolazioni pianificate utilizzate per lo stimatore”.

La voce “*Primo Dominio e analisi di controllo*” viene utilizzata per il calcolo dei pesi finali delle unità campionarie appartenenti alla prima partizione definita dalla variabile “Popolazioni pianificate utilizzate per lo stimatore”. Questa funzione è utile soprattutto per una prima analisi del problema: fornisce infatti le informazioni relative alla prima popolazione piani-

ficata e stampa in automatico alcune indicazioni che aiutano l'utente in una prima esplorazione dei risultati; tali dati sono anche ottenibili successivamente tramite la stampa 6 (*cfr. paragrafo 5.3*).

La voce “*Ulteriore Dominio*” può essere utilizzata solo se è stata precedentemente effettuata una elaborazione, attivata tramite una delle voci appena descritte e calcola i pesi finali considerando una specifica partizione scelta dall'utente. In tal caso però non effettua la stampa 3 “Statistiche sulle stime e i pesi diretti” e la stampa 6 “Tabelle di controllo” (*cfr. paragrafo 5.3*).

L'ultima voce “*Stampe*” permette all'utente di accedere alla maschera  $M_4$  direttamente, senza dover tornare alla schermata principale  $M_0$  per selezionare la funzione “Crea stampe”.

La voce “*Parametri*” – sia per la maschera  $M_2$  che per la maschera  $M_3$  – permette la selezione automatica delle variabili. Ciò è possibile solo se entrambi i *data-set* di input sono stati precedentemente utilizzati e si voglia fare uso delle medesime selezioni effettuate (nel successivo *paragrafo 5.2.3* viene descritto come utilizzare tale voce).

### **5.2.1 Le variabili e i parametri di input**

Il funzionamento del software prevede la definizione di alcune variabili di input e di alcuni parametri. Le **variabili dei data-set di input** corrispondono a quelle selezionabili tramite le maschere  $M_2$  ed  $M_3$  mostrate in *figura 5.5* e *figura 5.6* rispettivamente:

#### **Variabili del data-set “Totali Noti”:**

1. Popolazioni pianificate utilizzate per lo stimatore
2. Totali noti

#### **Variabili del data-set “Dati Campionari”:**

1. Popolazioni pianificate utilizzate per lo stimatore
2. Codice identificativo dell'unità campionaria
3. Variabili ausiliarie
4. Peso diretto
5. Peso distanza

La selezione delle variabili è obbligatoria, ad eccezione della variabile 5 (Peso distanza).

Tramite la maschera  $M_3$ , in cui si selezionano le variabili del *data-set* dei “Dati Campionari”, si effettua anche la scelta di due **parametri di input**. L’utente deve:

scegliere una **specifica popolazione pianificata** utilizzata per lo stimatore, nel caso in cui voglia calcolare i pesi finali con riferimento ad una sola popolazione.

scegliere una **funzione di distanza** e, ove necessario, i coefficienti moltiplicatori dei limiti dell’intervallo di variazione del peso finale.

La **costruzione dei data-set di input** per il calcolo dei pesi finali tramite la funzione di Riponderazione dipende dal disegno campionario utilizzato, dal tipo di stimatore, etc.

Essendo questa una operazione che esula dall’utilizzo vero e proprio delle maschere del software, si rimanda l’utente alla consultazione della *Sezione II*, avvertendolo che gli argomenti connessi con la costruzione dei *data-set* di input implicano una conoscenza approfondita delle scelte metodologiche alla base del campione in esame.

Nella *Sezione II* verrà anche approfondita la **trattazione dei parametri di input**.

### **5.2.2 La selezione delle variabili di input tramite le maschere di selezione**

La selezione di quasi tutte le variabili di input è obbligatoria per eseguire il calcolo dei pesi finali tramite le voci “Procedura Completa”, “Primo Dominio” e “Ulteriore Dominio” presentate nel paragrafo precedente. Se non sono inserite tutte le variabili obbligatorie, la procedura invia un messaggio di errore.

La selezione manuale delle variabili è descritta in questo paragrafo.

Come si può osservare dalle *figure 5.5 e 5.6* del precedente paragrafo, le maschere  $M_2$  e  $M_3$  permettono la selezione delle variabili: la prima

maschera serve a selezionare le variabili relative al *data-set* dei “totali noti”, la seconda per le variabili relative al *data-set* dei “dati campionari”.

L'utente deve selezionare le variabili per entrambi i *data-set*.

E' possibile elaborare i dati solo attivando la maschera  $M_3$  ed è perciò consigliabile iniziare la selezione delle variabili partendo dal *data-set* dei totali noti (maschera  $M_2$ ) per poi passare alla selezione delle variabili del *data-set* dei dati campionari.

Dopo la selezione delle variabili di un *data-set*, per passare alla selezione delle variabili dell'altro *data-set* è necessario tornare alla maschera  $M_1$  tramite la voce “Uscita” da “File”.

### Le variabili del *data-set* “Totali Noti”

La maschera  $M_2$  permette di selezionare le seguenti variabili:

- Popolazioni pianificate utilizzate per lo stimatore
- Totali noti

☐ Popolazioni pianificate utilizzate per lo stimatore  
☐ totali noti delle variabili  $X_1, \dots, X_n$

A titolo esemplificativo in *figura 5.5* viene mostrata la maschera che si attiva quando l'utente effettua la prima selezione.

**Figura 5.8 – Maschera di selezione della variabile “Popolazioni pianificate utilizzate per lo stimatore”.**



( $M_{21}$  attivata da  $M_2$ )

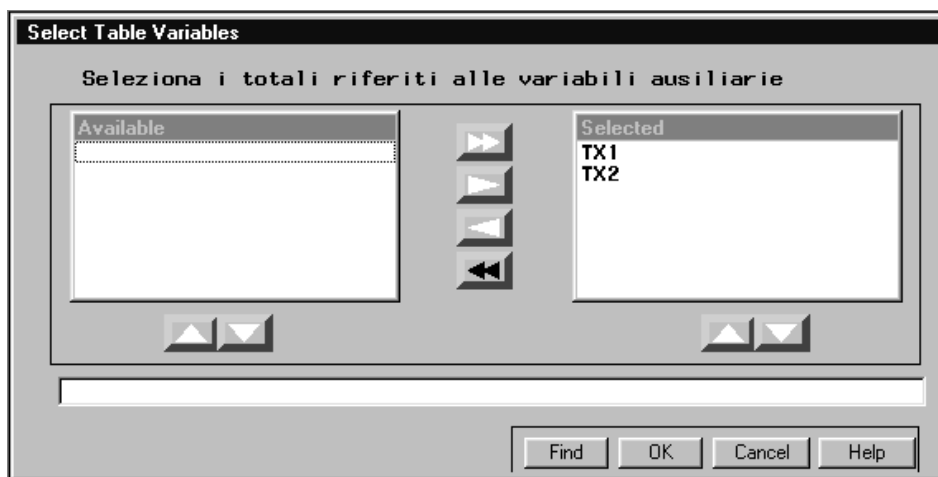




Le variabili possono essere selezionate singolarmente o in gruppo e spostate tramite la freccetta singola dal gruppo di sinistra (in cui vengono mostrate tutte le variabili disponibili – *available*) a quello di destra (in cui vengono poste le variabili selezionate – *selected*). I bottoni con le doppie freccette spostano tutte le variabili in entrambe le direzioni. Il tasto “Find” consente di trovare una determinata variabile tra quelle presenti nel *data-set* di input.

In *figura 5.8* è stata selezionata la variabile POPOL per definire le partizioni specificate dalla variabile “Popolazioni pianificate utilizzate per lo stimatore”; in *figura 5.9* sono stati selezionati i totali noti TX1 e TX2, relativi alle due variabili ausiliarie X1 ed X2, che verranno selezionate successivamente tramite la maschera  $M_3$  (cfr. *figura 5.10*). In questo caso sono possibili più selezioni. Quando la selezione è completa si utilizza il tasto “Ok” per tornare alla maschera  $M_2$ .

**Figura 5.9 – Maschera di selezione dei totali noti delle variabili ausiliarie.**



( $M_{22}$  attivata da  $M_2$ )

Terminate le scelte le variabili risultano selezionate.

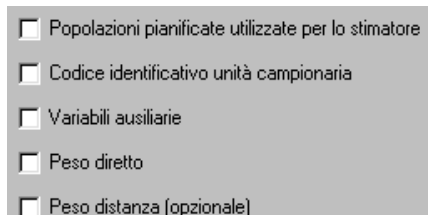
- ☒ Popolazioni pianificate utilizzate per lo stimatore
- ☒ totali noti delle variabili X1...Xn

## Le variabili del data-set “Dati Campionari”

Per selezionare le variabili del *data-set* dei dati campionari si utilizza la maschera  $M_3$  (cfr. figura 5.6).

Tramite questa maschera è possibile selezionare le variabili:


- Popolazioni pianificate utilizzate per lo stimatore
- Codice identificativo dell'unità campionaria
- Variabili ausiliarie
- Peso diretto
- Peso distanza



A vertical list of five checkboxes, all of which are currently unchecked. The labels are: "Popolazioni pianificate utilizzate per lo stimatore", "Codice identificativo unità campionaria", "Variabili ausiliarie", "Peso diretto", and "Peso distanza (opzionale)".

Tutte le selezioni effettuabili tramite la maschera  $M_3$  sono relative alla scelta di un'unica variabile, escludendo il caso delle variabili ausiliarie.

Titolo esemplificativo viene riportata in figura 5.10 la maschera di selezione delle variabili ausiliarie.



A single checkbox that is unchecked, followed by the text "Variabili ausiliarie".

Figura 5.10 – Maschera di selezione delle variabili ausiliarie.



The dialog box has a title bar "Select Table Variables". Inside, the main heading is "Seleziona le variabili ausiliare". It features two list boxes: "Available" on the left containing "COEF", "CK", and "CODICE"; and "Selected" on the right containing "X1" and "X2". Between the boxes are four arrow buttons for moving items. Below the boxes are two small up/down arrow buttons. At the bottom right are four buttons: "Find", "OK", "Cancel", and "Help".

( $M_{31}$  attivata da  $M_3$ )

## I parametri di input per l'utilizzo del software

Tramite la maschera  $M_3$  è anche possibile scegliere alcuni parametri fondamentali per il funzionamento del software.

**Figura 5.11 – La selezione dei parametri di input**

☒ Popolazioni pianificate utilizzate per lo stimatore  
☒ Codice identificativo unità campionaria  
☒ Variabili ausiliarie  
☒ Peso diretto  
☒ Peso distanza (opzionale)

Funzioni di distanza  
Logaritmica troncata  
Logaritmica  
Euclidea  
Lineare troncata  
Hellinger  
Minima Entropia  
Chi quadrato

0.5 Coeff. moltiplicatore valore minimo di L  
1.5 Coeff. moltiplicatore valore massimo di U

Pop. pianif. per stimatore  
18  
19  
20

2 Riga selezionata

In particolare, l'utente può:

- come **primo parametro**: scegliere una **specificata popolazione pianificata** utilizzata per lo stimatore, nel caso utilizzi la voce "Ulteriore Dominio". Come visto nel *paragrafo* 5.2, la voce "Ulteriore Dominio" è disponibile per il calcolo dei pesi finali solo se vi è stata una precedente elaborazione. In tal caso appare la lista dei codici della variabile "Popolazioni pianificate utilizzate per lo stimatore" e l'utente può selezionare la partizione di interesse (nella figura, la selezione in oggetto si riferisce alla partizione "19" che corri-

Pop. pianif. per stimatore  
18  
19  
20

2 Riga selezionata

sponde alla riga “2”). Il software in questo caso, calcola i pesi finali rispetto alla partizione selezionata.

- b) come **secondo parametro**: scegliere una **funzione di distanza** e, ove necessario, i **coefficienti moltiplicatori dei limiti dell’intervallo di variazione del peso finale**.

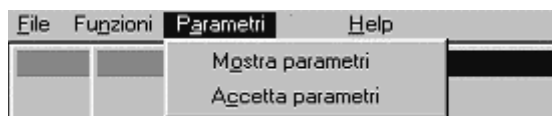
Come mostrato in *figura 5.6* il software permette di considerare sette possibili tipi di distanza

Per maggiori dettagli sulla selezione di questi parametri, si legga il *paragrafo 1.1.2*, *Sezione II*, in cui, in particolare, si approfondisce l’uso del secondo parametro, anche in relazione alla convergenza della procedura iterativa sottostante.

### 5.2.3 La selezione delle variabili di input tramite i parametri attivati dal software

Come visto nel *paragrafo 5.2.2* le variabili possono essere selezionate tramite le relative maschere  $M_2$  ed  $M_3$ . Esiste una alternativa per agevolare l’utente: la funzione “Parametri”, mostrata in *figura 5.12*.

**Figura 5.12.** I “Parametri” attivati dalle maschere  $M_2$  ed  $M_3$



Tale funzione è utilizzabile solo se è stata effettuata una precedente elaborazione con gli stessi *data-set* di input (in altri termini è stata precedentemente attivata una delle voci “Procedura Completa” o “Primo Dominio” o “Ulteriore Dominio”), in quanto il software crea nella cartella di output il *data-set* SAVETIME.sas7bdat che memorizza i parametri della elaborazione (cfr. *paragrafo 2.1*, *Sezione II*).

Per usufruire di tale possibilità, nella elaborazione successiva occorre scegliere le stesse cartelle di input ed output della elaborazione precedente (ovviamente scegliendo la stessa cartella di output in diverse elaborazioni, il programma sovrascrive i *data-set* precedentemente memorizzati).

Come mostrato in *figura 5.12*, l'utente può scegliere le due voci “Mostra parametri” o “Accetta parametri”.

La prima voce permette la visualizzazione della maschera di *figura 5.13*, tramite la quale l'utente può visualizzare i parametri.

Successivamente, per accettare i parametri visualizzati, l'utente deve scegliere la voce “Accetta parametri”, rendendo in tal modo attive tutte le variabili utilizzate nella precedente elaborazione, anche per lo specifico *data-set* su cui sta operando.

In particolare, se si utilizza la maschera  $M_2$ , risulteranno selezionate automaticamente le variabili del *data-set* dei totali noti, al contrario se si utilizza la maschera  $M_3$ , verranno recepite le scelte del *data-set* dei dati campionari. Non è perciò possibile selezionare le variabili per entrambi i *data-set*. E' invece comunque possibile modificare manualmente la scelta.

**Figura 5.13: Parametri attivi**


Parametri utilizzati per calcolo coeff. riporto all'universo		
	descr	parametro
1	INPUT1 - dsn dati campionari	C:\STIME INP\Inptime.Input
2	INPUT2 - dsn totali	C:\STIME INP\Inptime.Noti
3	Peso distanza	PESO_DIST
4	Variabili ausiliarie	X1 X2
5	Peso diretto	PESO_INIZ
6	Pop. pianif. per stimatore INPUT1	REGIONE
7	Codice identif. unità campionaria	CODICE
8	Cod. pop. pianif. per stimatore	.
9	Pop. pianif. per stimatore INPUT2	REGIONE
10	Totali variabili ausiliarie	TX1 TX2
11	Funzione di distanza	1
12	Coeff.molt.val.minimo di L	0.5
13	Coeff.molt.val.massimo di U	1.5

( $M_{32}$  attivata da  $M_3$ )

In *figura 5.13* il parametro relativo alla riga 8 (Cod.pop.pianif.per stimatore) è un valore mancante, in quanto l'elaborazione è stata attivata tramite “Procedura completa” o “Primo dominio e analisi di controllo”. Nel caso si fosse utilizzata la voce “Ulteriore dominio”, in tale riga apparirebbe il

numero di riga corrispondente alla partizione selezionata (cfr. figura 5.11 Sezione I)

### 5.3 La funzione “crea stampe”

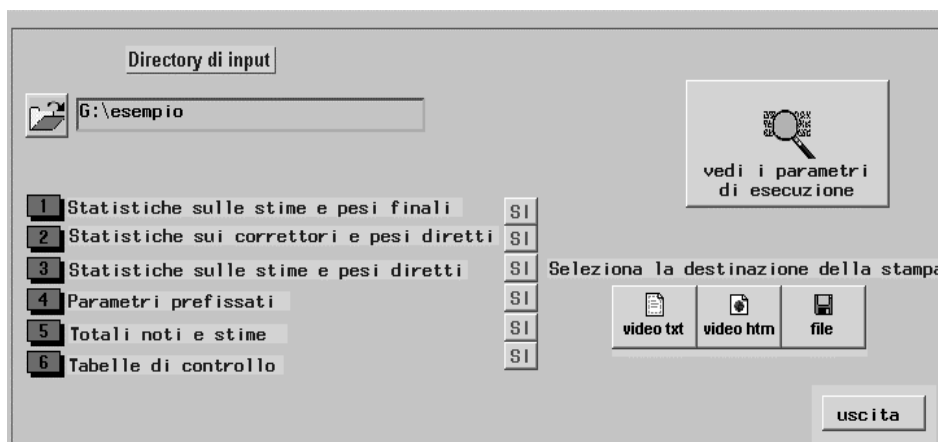
La funzione  esegue le stampe riferite ad elaborazioni precedenti. Essa attiva la schermata di figura 5.14. La cartella di input per le stampe corrisponde alla cartella che contiene i *data-set* di output di una precedente elaborazione.

Una volta selezionata la cartella, si scelgono le stampe desiderate tramite i bottoni dove appare il “SI”, valore che appare per *default* e che può essere variato.



Se ad esempio si volesse ottenere solo la stampa numero 1 “Stime e pesi finali”, si dovrà fare in modo che appaia il “SI” sul bottone relativo alla prima stampa, mentre per le altre stampe si varia il valore di *default* da “SI” a “NO”.

**Figura 5.14 – Maschera di selezione delle stampe**



( $M_4$  attivata da  $M_1$ )

Come si vede in figura 5.14, è possibile elaborare 6 differenti stampe; l’ultima – “Tabelle di controllo” – elabora diversi tabulati di analisi. Essendo possibile elaborare le stampe in un momento successivo a quello in cui

sono stati creati i *data-set* di output, è sempre consentita la visione dei parametri di esecuzione della procedura.

Si può scegliere di effettuare, con appositi bottoni, le stampe a video in formato html<sup>2</sup> o txt; in questo secondo caso, è possibile visualizzare le stampe tramite l'opzione "output", che si attiva nel menu bar. In alternativa è possibile produrre dei file che memorizzano le stampe: stampa1.txt, stampa2.txt, stampa3.txt, stampa4.txt, stampa5.txt e stampa6.txt.

In questo paragrafo, a titolo di esempio, vengono mostrate le stampe che si ottengono quando l'utente sceglie la visualizzazione a video (opzione "video.htm").

Per approfondimenti sulle stampe si legga il *paragrafo 5.4* in cui vengono dettagliatamente descritti i campi presenti nelle maschere, seguendo la stessa numerazione progressiva riportata in questo paragrafo.

La prima stampa è presentata in *figura 5.15*. La *tavola 1* riporta le statistiche sulle stime e sui pesi finali per ciascun codice di popolazione pianificata utilizzata per lo stimatore.

**Figura 5.15: Tavola 1- Statistiche sulle stime e i pesi finali**

**Tav.1: Statistiche sulle stime e pesi finali per pop. pianif. stimatore**

Pop. pianif. stimatore	Minimo peso finale	Massimo peso finale	Massima differenza stime	Minima differenza stime	Somma dei pesi finali	Media dei pesi finali	Coeff. var. (% dei pesi finali)
18	1.92249	704.457	-0.00	-0.00	48408.00	48.46	164.1
19	2.19282	873.916	-0.00	-0.00	24803.00	38.57	200.9
20	1.80831	530.301	0.00	-0.00	50207.00	48.89	139.4
					<b>123418.00</b>		

<sup>2</sup>

Si noti che le stampe a video in formato html, una volta prodotte, vengono poste in secondo piano rispetto alla maschera attiva. Per portarle in primo piano, è sufficiente cliccare con il mouse dove è visibile la stampa. Per abbandonare le stampe e tornare alla procedura si può usare il tasto PF3

In particolare presenta i valori relativi ai campi:

- Popolazioni pianificate utilizzate per lo stimatore
- Minimo peso finale
- Massimo peso finale
- Massima differenza stime
- Minima differenza stime
- Somma dei pesi finali
- Media dei pesi finali
- Coeff. Var. (%) dei pesi finali

Una descrizione dettagliata di tali statistiche è riportata nel *paragrafo 6.1* del prossimo capitolo.

La seconda stampa è presentata in *figura 5.16*. Riporta alcune statistiche relative ai valori assunti dai correttori per ciascun codice di popolazione pianificata utilizzata per lo stimatore.

**Figura 5.16: - Tavola 2 – Statistiche sui correttori dei pesi diretti**

**Tav. 2: Statistiche sui correttori per pop. pianif. stimatore**

Pop. pianif. stimatore	L_MAX	U_MIN	L_EFF	U_EFF	Minimo correttore osservato	Massimo correttore osservato	Somma dei correttori	Media dei correttori	Coeff. var. (%) dei correttori
18	1.00	1.11	0.80000	1.00900	1.01940	1.21984	1098.70	1.10	6.56
19	1.00	1.13	0.80000	1.76328	0.94033	1.73385	725.22	1.13	20.35
20	0.95	1.11	0.47522	1.36461	0.84205	1.42203	1072.92	1.04	17.18
							<b>2896.84</b>		

In particolare presenta i valori relativi ai campi:

9. Pop. Pianif. Stimatore
10. L\_MAX
11. U\_MIN
12. L\_EFF
13. U\_EFF
14. Minimo correttore osservato



15. Massimo correttore osservato
16. Somma dei correttori
17. Media dei correttori
18. Coeff. Var. (%) dei correttori

Una descrizione dettagliata di tali statistiche è riportata nel *paragrafo 6.2* del prossimo capitolo.

La terza stampa è presentata in *figura 5.17*. La *tavola 3* riporta le statistiche sulle stime e sui pesi diretti per ciascun codice di popolazione pianificata utilizzata per lo stimatore.

**Figura 5.17: Tavola 3 – Statistiche sulle stime e pesi diretti**

<b>Tav.3: Statistiche sulle stime e sui pesi diretti per pop. pianif. stimatore</b>							
<b>Pop. pianif. stimatore</b>	<b>Minimo peso diretto</b>	<b>Massimo peso diretto</b>	<b>Massima differenza stime</b>	<b>Minima differenza stime</b>	<b>Somma dei pesi diretti</b>	<b>Media dei pesi diretti</b>	<b>Coeff. var. (%) dei pesi diretti</b>
<b>18</b>	1.67607	588.254	-4903.36	-12430.58	43504.64	43.55	159.87
<b>19</b>	2.31800	679.857	-2120.65	-3703.56	21099.64	32.81	190.84
<b>20</b>	2.14599	417.036	8440.87	-4970.54	45236.66	44.05	134.32
					<b>109840.94</b>		

In particolare presenta i valori relativi ai campi:

19. Pop. Pianif. Stimatore
20. Minimo peso diretto
21. Massimo peso diretto
22. Massima differenza stime
23. Minima differenza stime
24. Somma dei pesi diretti
25. Media dei pesi diretti
26. Coeff. Var. (%) dei pesi diretti

Una descrizione dettagliata di tali statistiche è riportata nel *paragrafo 6.3* del prossimo capitolo.

La quarta stampa è presentata in *figura 5.18* e mostra alcune informazioni relative alle procedura iterativa per il calcolo dei pesi finali per ciascun codice di popolazione pianificata utilizzata per lo stimatore.

**Figura 5.18: Tavola 4 – Parametri prefissati per la procedura iterativa di stima**

**Tav. 4: Parametri prefissati per la procedura iterativa di stima**

Pop. pianif. stimatore	Numero di iterazioni effettuate	Numero di stime vincolate	Numero di unità rilevate	Numero massimo di iterazioni	Funzione di distanza
<b>18</b>	4	2	999	30	LGT
<b>19</b>	5	2	643	30	LGT
<b>20</b>	5	2	1027	30	LGT
			<b>2669</b>		

In particolare presenta i valori relativi ai campi:

27. Pop. Pianif. Stimatore
28. Numero di iterazioni effettuate
29. Numero di stime vincolate
30. Numero di unità rilevate
31. Numero massimo di iterazioni
32. Funzione di distanza

Una descrizione dettagliata di tali statistiche è riportata nel *paragrafo 6.4* del prossimo capitolo.

La quinta stampa è presentata in *figura 5.19* e riporta diverse informazioni relative a totali noti, stime finali e differenze per ciascun codice di popolazione pianificata utilizzata per lo stimatore.

**Figura 5.19: Tavola 5 – Totali noti, stime dirette, finali e differenze**

**Tav.5: Totali noti, stime dirette, stime finali e differenze**

Pop. pianif. stimatore	Codice di totale	Totali noti	Stime finali	Stime dirette	Differenza stime finali	Differenza stime dirette	Totali campionari
18	1	18408	18408	13505	0	1903	999.0
18	2	27374	27374	261314	-0	-12430	10207.9
19	1	24803	24803	21100	-0	-3703	643.0
19	2	185610	185610	183789	0	2121	11082.7
20	1	50207	50207	45237	0	4970	1027.0
20	2	161856	161856	170297	-0	8441	10933.8
		<b>744628</b>	<b>744628</b>	<b>724941</b>	<b>-0</b>	<b>-19687</b>	<b>34893.1</b>

In particolare presenta i valori relativi ai campi:

33. Pop. Pianif. Stimatore
34. Codice di totale
35. Totali noti
36. Stime finali
37. Stime dirette
38. Differenza stime finali
39. Differenza stime dirette
40. Totali campionari

Una descrizione dettagliata di tali statistiche è riportata nel *paragrafo 6.5* del prossimo capitolo.

I tre tabulati ottenibili tramite la *stampa 6* sono presentati nelle *figure 5.20 – 5.22*.

Una descrizione dettagliata di tutti i tabulati è riportata nel *paragrafo 6.6* del prossimo capitolo.

**Figura 5.20: Tabulato 1 – Controllo sulle celle per dominio: Totali noti, stime dirette, Rapporti tra totali noti e stime dirette, Totali campionari**

**Tab.1: Controllo sulle celle per pop. pianif. stimatore: Totali noti, Stime dirette, Rapporti tra Totali noti e Stime dirette, Totali campionari**

**nr.progr. popolaz.=1**

<b>Pop. pianif. stimatore</b>	<b>Codici di totale</b>	<b>Totali Noti</b>	<b>Stime Dirette</b>	<b>Rapporto Totale noto e Stima diretta</b>	<b>Totali Campionari</b>
<b>18</b>	1	43403	43504.64	1.11	999.0
<b>18</b>	2	273744	261313.62	1.05	10207.9
<b>18</b>	TOTALE	322152	304818.26	1.06	11206.9

**nr.progr. popolaz.=2**

<b>Pop. pianif. stimatore</b>	<b>Codici di totale</b>	<b>Totali Noti</b>	<b>Stime Dirette</b>	<b>Rapporto Totale noto e Stima diretta</b>	<b>Totali Campionari</b>
<b>19</b>	1	24803	21099.84	1.18	543.0
<b>19</b>	2	185310	183489.35	1.01	11082.4
<b>19</b>	TOTALE	210113	204589.19	1.03	11625.4

Il tabulato 1 presenta diverse sottotabelle identificate dal un numero progressivo per ciascun codice di popolazione pianificata per lo stimatore e presenta inoltre i seguenti campi:

41. Pop. Pianif. Stimatore
42. Codice di totale
43. Totali noti
44. Stime dirette
45. Rapporto Totale noto e Stima diretta
46. Totali campionari

**Figura 5.21: Tabulato 2 – Campione dei rispondenti e stima del totale popolazione**

**Tab.2: Campione rispondenti e stima della popolazione con i pesi diretti**

Pop. pianif. stimatore	nr.progr. popolaz.	Campione rispondenti (nr. unità rilevate)	Somma dei pesi diretti
<b>18</b>	1	999	43504.64
<b>19</b>	2	643	21099.64
<b>20</b>	3	1027	45236.66
<b>TOTALE</b>	4	2669	109840.94

Il tabulato 2 presenta i seguenti campi:

- 47. Pop. Pianif. Stimatore
- 48. nr. progr. popolaz.
- 49. Campione rispondenti (nr. unità rilevate)
- 50. Somma dei pesi diretti

L'ultimo tabulato appare solo se il *data-set* dei "Totali Noti" si riferisce ad una o più popolazioni pianificate che non trovano riscontro nel *data-set* dei "Dati Campionari". Il software crea il *tabulato 3* che presenta le popolazioni pianificate che non hanno dati campionari (ovvero possibili mancate risposte) e scrive le corrispondenti informazioni nel *data-set* VUOTI (cfr. capitolo 2, Sezione II).

---

**Figura 5.22: Tabulato 3 – Controllo sulle popolazioni pianificate che non hanno dati campionari**

**Tab.3: Controllo su pop. pianif. che non hanno dati campionari**

<b>Pop. pianif. stimatore</b>	<b>totale_noto_variabile_1</b>	<b>totale_noto_variabile_2</b>
<b>45</b>	3444	6666

---

Il tabulato 3 presenta i seguenti campi:

- 51. Pop. Pianif. Stimatore
- 52. Totale\_noto\_variabile\_1
- 53. Totale\_noto\_variabile\_2



## 6. La descrizione delle stampe prodotte dalla funzione di Riponderazione del software Genesees V. 3.0

Il software oltre a calcolare i coefficienti finali di riporto, produce alcune statistiche descrittive di sintesi, presentate attraverso tavole e tabulati, relative al processo di ponderazione vincolata che ha generato i coefficienti finali di riporto.

### Avvertenze per una migliore lettura

Nei paragrafi che seguono vengono descritti tutti i campi delle stampe che sono mostrate nelle diverse figure del *paragrafo 5.3* (*cfr. figure 5.15-5.22, capitolo 5*); tali campi seguono la stessa numerazione progressiva, in modo da identificare facilmente la stampa a cui si riferiscono. E' inoltre da evidenziare che nel seguito – oltre che al suddetto *paragrafo 5.3* - si fa spesso riferimento anche al *paragrafo 1.2* della *Sezione II*; è dunque consigliabile prendere visione anche di tale paragrafo.

### 6.1 Tavola 1 – Statistiche sulle stime e pesi finali per popolazione Pianif. Stimatore

La tavola presenta alcune statistiche relative ai pesi o coefficienti finali di riporto e alle stime ottenute con questi coefficienti (*cfr. figura 5.15, capitolo precedente*). Queste si riferiscono ai sottocampioni che appartengono alla stessa sottopopolazione pianificata e sono calcolate sui record che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** (indicata sinteticamente come POP\_PIAN nel *paragrafo 1.2, Sezione II*).



Le variabili e le statistiche della tavola sono:

1. **Pop. Pianif. Stimatore:** indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Tale variabile identifica le righe della tavola;
2. **Minimo peso finale:** indica il coefficiente finale di riporto minimo, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
3. **Massimo peso finale:** indica il coefficiente finale di riporto massimo, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
4. **Massima differenza stime:** indica la massima differenza tra le stime dei totali delle variabili ausiliarie, ottenute con i coefficienti finali di riporto, e i rispettivi totali noti. Tale valore è ottenuto considerando tutti i valori delle variabili **Totali noti** (indicate sinteticamente nel *capitolo 4* con TXj) e le rispettive stime. Le differenze sono calcolate per ciascun sottocampione che presenta la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
5. **Minima differenza stime:** indica la minima differenza tra le stime dei totali delle variabili ausiliarie, ottenute con i coefficienti finali di riporto, e i rispettivi totali noti. Tale valore è ottenuto considerando tutti i valori delle variabili **Totali noti** (indicate sinteticamente nel *paragrafo 1.2, Sezione II* con TXj) e le rispettive stime. Le differenze sono calcolate per ciascun sottocampione che presenta la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
6. **Somma dei pesi finali:** indica la somma dei coefficienti finali di riporto. Tale somma rappresenta una stima della sottopopolazione che presenta la modalità indicata daella variabile **Popolazioni pianificate utilizzate per lo stimatore**. In coda a tutti i valori viene presentata la somma complessiva, che rappresenta la stima della dimensione della popolazione obiettivo secondo lo stimatore di ponderazione vincolata adottato;

7. **Media dei pesi finali:** indica il coefficiente finale di riporto medio, registrato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
8. **Coeff. Var. (%) dei pesi finali:** indica il coefficiente di variazione percentuale dei coefficienti finali di riporto, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola.

## 6.2 Tavola 2 – Statistiche sui correttori per popolazione Pianif. Stimatore

La tavola presenta alcune statistiche relative ai correttori (generati dal processo di calibrazione) che trasformano i pesi o coefficienti iniziali (diretti) di riporto nei pesi o coefficienti finali (cfr. *figura 5.16, capitolo precedente*). Tali statistiche sono calcolate sui sottocampioni che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** (POP\_PIAN). Le variabili della tavola sono:

9. **Pop. Pianif. Stimatore:** indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Tale variabile identifica le righe della tavola;
10. **L\_MAX:** limite inferiore teorico dell'intervallo di variazione dei correttori,  $L_{MAX}$ , (cfr. *capitolo 4*). Il limite inferiore  $L_{max}$  è ottenuto considerando i rapporti  $R_t = \frac{X_t}{\sum_{k \in S} x_{tk} d_k}$ , in cui il numeratore è il totale noto della variabile ausiliaria  $x_t$  e il denominatore è la stima del totale calcolata con i pesi diretti sul sottocampione individuato dalla modalità della prima variabile della tavola.  $L_{max}$  è pari al minimo valore degli  $R_t$ . Se  $R_t \geq 1$  per ogni  $t$  ( $t=1, \dots, T$ ) si pone  $L_{max} = 0,99$ ;
11. **U\_MIN:** limite superiore teorico dell'intervallo di variazione dei correttori,  $U_{min}$ , (cfr. *capitolo 4*). Il limite superiore  $U_{min}$  è ottenu-

to considerando i rapporti  $R_t = \frac{X_t}{\sum_{k \in S} x_{tk} d_k}$ , in cui il numeratore è

il totale noto della variabile ausiliaria  $x_t$  e il denominatore è la stima del totale calcolata con i pesi diretti sul sottocampione individuato dalla modalità della prima variabile del tavolo.  $U_{\min}$  è pari al massimo valore degli  $R_t$ . Se  $R_t \leq 1$  per ogni  $t$  ( $t=1, \dots, T$ ) si pone  $U_{\min} = 1,01$ ;

12. **L\_EFF**: indica il prodotto tra **L\_MAX** e il coefficiente moltiplicatore (per default pari a 0,5), richiamato nelle maschere di lancio della procedura come *Coeff. Moltiplicatore valore minimo di L* (per approfondimenti cfr. capitolo 4). Tale statistica ha valore quando è stata adottata una funzione di distanza troncata per la calibrazione dei coefficienti finali di riporto;
13. **U\_EFF**: indica il prodotto tra **U\_MIN** e il coefficiente moltiplicatore (per default pari a 1,5), richiamato nelle maschere di lancio della procedura come *Coeff. Moltiplicatore valore massimo di U* (per approfondimenti cfr. capitolo 4). Tale statistica ha valore quando è stata adottata una funzione di distanza troncata per la calibrazione dei coefficienti finali di riporto;
14. **Minimo correttore osservato**: indica il correttore minimo dei coefficienti diretti calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
15. **Massimo correttore osservato**: indica il correttore massimo del coefficiente diretto calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
16. **Somma dei correttori**: indica la somma dei correttori dei coefficienti iniziali nel sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore**. In coda a tutti i valori viene presentata la somma

complessiva dei correttori calcolata sull'intero campione dei rispondenti;

17. **Media dei correttori:** indica il correttore medio dei coefficienti diretti calcolati all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
18. **Coeff. Var. (%) dei correttori:** indica il coefficiente di variazione percentuale dei correttori calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola.

### 6.3 Tavola 3 – Statistiche sulle stime e sui pesi diretti per popolazione Pianif. Stimatore

La tavola presenta alcune statistiche relative ai coefficienti iniziali di riporto e alle stime ottenute con questi coefficienti (cfr. *figura 5.17, capitolo precedente*). Queste sono calcolate considerando i sottocampioni di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** (POP\_PIAN). Le variabili della tavola sono:

19. **Pop. Pianif. Stimatore:** indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Tale variabile identifica le righe della tavola;
20. **Minimo peso diretto:** indica il coefficiente diretto di riporto minimo, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
21. **Massimo peso diretto:** indica il coefficiente diretto di riporto massimo, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;

22. **Massima differenza stime:** indica la massima differenza tra la stima dei totali delle variabili ausiliarie, ottenuta con i coefficienti iniziali di riporto, e i rispettivi totali noti utilizzati nella calibrazione. Tale valore è ottenuto considerando tutti i valori delle variabili **Totali noti** (indicate sinteticamente nel *paragrafo 1.2.1, Sezione II* con TXj) del *data-set “totali noti”* e le rispettive stime. Le differenze sono calcolate per ciascun sottocampione che presenta la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
23. **Minima differenza stime:** indica la minima differenza tra le stime dei totali delle variabili ausiliarie, ottenute con i coefficienti iniziali di riporto, e i rispettivi totali noti utilizzati nella calibrazione. Tale valore è ottenuto considerando tutti i valori delle variabili **Totali noti** (TXj) del *data-set “totali noti”* e le rispettive stime. Le differenze sono calcolate per ciascun sottocampione che presenta la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
24. **Somma dei pesi diretti:** indica la somma dei coefficienti diretti di riporto. Tale somma rappresenta una stima della sottopopolazione (con i pesi diretti) che presenta la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola. In coda a tutti i valori viene presentata la somma complessiva che rappresenta la stima della dimensione della popolazione di riferimento secondo lo stimatore diretto (Horvitz-Thompson) adottato;
25. **Media dei pesi diretti:** indica il coefficiente iniziale di riporto medio registrato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola;
26. **Coeff. Var. (%) dei pesi diretti:** indica il coefficiente di variazione percentuale dei coefficienti diretti di riporto, calcolato all'interno del sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola.

#### 6.4 Tavola 4 – Parametri prefissati per la procedura iterativa di stima

In questa tavola sono fornite alcune principali caratteristiche del processo di ponderazione vincolata che è stato adottato dall'utente per ciascun sottocampione identificato dalle unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** (indicata sinteticamente come POP\_PIAN nel *paragrafo 1.2, Sezione II* (cfr. *figura 5.18, capitolo precedente*)). Le variabili della tavola sono:

27. **Pop. Pianif. Stimatore:** indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Tale variabile identifica le righe della tavola;
28. **Numero di iterazioni effettuate:** indica il numero di iterazioni effettuate per ottenere la convergenza delle stime dei totali, calcolate con i coefficienti finali di riporto, ai rispettivi totali noti. Tali totali sono stati indicati sinteticamente nel *paragrafo 1.2, Sezione II*, con le variabili TXj. Per quanto riguarda gli stimatori che utilizzano la distanza Euclidea o lineare, il numero di iterazioni è sempre pari a zero, in quanto esiste una soluzione esplicita al problema di calibrazione e quindi non è necessario procedere con metodi iterativi (per approfondimenti cfr. *capitolo 4*);
29. **Numero di stime vincolate:** indica il numero dei totali noti utilizzati nel processo di calibrazione per il sottocampione di unità che presentano la stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola (tale valore indica il numero delle variabili chiamate TXj *paragrafo 1.2, , Sezione II*);
30. **Numero di unità rilevate:** numero di unità elementari rispondenti (record) presenti nel sottocampione individuato dalla modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** indicata nella prima colonna della tavola. In coda a tutti i valori viene presentata la somma complessiva del numero di unità rilevate pari alla dimensione del campione dei rispondenti;
31. **Numero massimo di iterazioni;** indica il massimo numero di iterazioni per ottenere la convergenza delle stime dei coefficienti

finali con i relativi totali noti. Per default il numero massimo di iterazioni è posto pari a 30;

32. **Funzione di distanza:** indica la funzione di distanza utilizzata. Le sigle che compaiono corrispondono alle seguenti distanze: LGT – Logaritmica troncata; LOG – Logaritmica; LIN – Euclidea (o Lineare); LIT – Lineare troncata; HEL – Hellinger; ENT – Minima Entropia; CHI – Chi Quadrato.

## 6.5 Tavola 5 – Totali noti, stime dirette, stime finali e differenze

La tavola presenta alcune informazioni relative alle caratteristiche dello stimatore adottato dall'utente (cfr. *figura 5.19, capitolo precedente*). Tra queste variabili, molto importante risulta quella indicata come **Differenza stime finali**, attraverso la quale è possibile verificare se i coefficienti finali riproducono i totali noti (in questa circostanza la variabile è sempre pari a zero) oppure se tale obiettivo non è stato raggiunto (qualche valore della variabile è diverso da zero).

Qualora si presenti questa seconda eventualità, per ottenere l'uguaglianza tra le stime delle variabili ausiliarie ottenute con i coefficienti finali e i relativi totali noti, è necessario distinguere due casi: il primo è relativo agli stimatori che presentano una soluzione analitica del problema (stimatori che utilizzano la distanza Euclidea); il secondo è relativo agli stimatori che trovano una soluzione attraverso metodi iterativi (stimatori che non utilizzano la distanza Euclidea). Per la prima classe di stimatori il problema sottostante alla ponderazione vincolata deve essere riformulato, in quanto il numero delle osservazioni utilizzate nel sistema di calibrazione dei coefficienti non è sufficientemente ampio (per approfondimenti cfr. *capitolo 4 e l'appendice A.1*).

Nel secondo caso, la non convergenza potrebbe invece dipendere anche da altri fattori, quali, ad esempio: un numero di iterazioni non sufficientemente ampio affinché le stime convergano ai totali noti; un intervallo troppo piccolo tra il valore del *Coeff. Moltiplicativo valore minimo di L* e il valore del *Coeff. Moltiplicativo valore massimo di U* quando si utilizzano funzioni di distanza troncate.

Nella tavola, ciascuna riga si riferisce ai sottocampioni che presentano la

stessa modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** e ad un particolare totale noto utilizzato per la calibrazione. La tavola presenta le seguenti variabili:

33. **Pop. Pianif. Stimatore:** indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Ciascuna modalità è replicata per il numero dei totali noti considerati;
34. **Codice di totale:** la variabile assume una numerazione progressiva che segue l'ordine con cui sono state inserite le variabili ausiliarie al momento del lancio della procedura. Le informazioni successive della tavola fanno riferimento alla sottopopolazione identificata dalla modalità della prima colonna della tavola e al rispettivo totale indicato da questa variabile;
35. **Totali noti:** indica il totale noto sulla data sottopopolazione, identificato dalla modalità della prima colonna della tavola relativamente alla variabile indicata dalla colonna **Codice di totale**. In coda a tutti i valori viene presentato il totale complessivo per tutte le variabili ausiliarie;
36. **Stime finali:** stima ottenuta con i coefficienti finali di riporto del totale indicato dalla colonna **Codice di totale** per la sottopopolazione definita dalla modalità della prima colonna della tavola. In coda a tutti i valori viene presentato il totale complessivo stimato con i coefficienti finali per tutte le variabili ausiliarie;
37. **Stime dirette:** stima ottenuta con i coefficienti iniziali o diretti di riporto del totale indicato dalla colonna **Codice di totale** per la sottopopolazione definita dalla modalità della prima colonna della tavola. In coda a tutti i valori viene presentato il totale complessivo stimato con i coefficienti iniziali per tutte le variabili ausiliarie;
38. **Differenza stime finali:** differenza tra la stima ottenuta con i coefficienti finali e il totale noto della variabile ausiliaria utilizzata nella calibrazione ed indicata nella colonna **Codice di totale**. Tale differenza è calcolata sul sottocampione indicato dalla modalità della colonna **Pop. Pianif. Stimatore**. In coda a tutti i valori viene presentato il totale complessivo delle differenze tra totale stimato con i coefficienti finali e totali noti per tutte le variabili ausiliarie;



39. **Differenza stime dirette:** differenza tra la stima ottenuta con i coefficienti iniziali o diretti e il totale noto della variabile ausiliaria utilizzata nella calibrazione ed indicata nella colonna **Codice di totale**. Tale differenza è calcolata sul sottocampione indicato dalla modalità della colonna **Pop. Pianif. Stimatore**. In coda a tutti i valori viene presentato il totale complessivo delle differenze tra totale stimato con i coefficienti iniziali e totali noti per tutte le variabili ausiliarie;
40. **Totali campionari:** totale ottenuto senza utilizzare i coefficienti di riporto nei sottocampioni identificati dalla modalità della prima colonna della tavola relativamente alla variabile ausiliaria individuata dalla colonna **Codice di totale**. In coda a tutti i valori viene presentato il totale complessivo delle stime senza l'uso dei coefficienti di riporto dei totali noti per tutte le variabili ausiliarie.

## 6.6. Tabulati di controllo

Attraverso la stampa 6 “Tabelle di controllo”, il software produce automaticamente 3 tabulati.

### Tab. 1: Controllo sulle celle per pop. pianif. stimatore: Totali noti, Stime dirette, Rapporti tra Totali noti e Stime dirette, Totali campionari

Il tabulato 1 presenta alcune statistiche sul disegno di campionamento e lo stimatore iniziale, considerando singolarmente ciascun sottocampione identificato da una particolare modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore** (cfr. *figura 5.20, capitolo precedente*). Il tabulato è costituito da sottotabelle, ciascuna delle quali è identificata da una modalità della variabile **nr. progr. popolaz..** Tale variabile rinumerale le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. Le variabili contenute sono le seguenti:

41. **Pop. Pianif. stimatore:** indica la modalità corrispondente alla ricodifica di **nr. progr. popolaz.** con valori assunti dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**;
42. **Codice di totale:** la variabile assume una numerazione progressi-

va, che segue l'ordine con cui sono state inserite le variabili ausiliarie al momento del lancio della procedura. Le informazioni successive del tabulato fanno riferimento alla modalità assunta da tale variabile. L'ultima riga, denominata **TOTALE**, indica che le statistiche successive sono calcolate considerando tutti i totali noti;

43. **Totali noti**: indica il totale noto della sottopopolazione, identificata dalla modalità della prima colonna del tabulato, relativamente alla variabile indicata dalla colonna **Codice di totale**; L'ultima riga denominata **TOTALE** è la somma delle righe precedenti;
44. **Stime dirette**: stima ottenuta con i coefficienti iniziali di riporto del totale indicato dalla colonna **Codice di totale** per la sottopopolazione definita dalla modalità della prima colonna del tabulato. L'ultima riga, denominata **TOTALE**, è la somma delle righe precedenti;
45. **Rapporto Totale noto e Stima diretta**: indica il rapporto tra i valori delle variabili **Totali noti** e **Stime dirette**;
46. **Totali campionari**: totale ottenuto, senza utilizzare i coefficienti di riporto, nei sottocampioni identificati dalla modalità della prima colonna della tavola, relativamente alla variabile ausiliaria individuata dalla colonna **Codice di totale**. L'ultima riga, denominata **TOTALE**, è la somma delle righe precedenti.

L'ultima sottotabella del tabulato 1 si riferisce a tutto il campione. In questo caso la colonna **Pop. Pianif. stimatore** assume la modalità **TOTALE**.

## **Tab. 2: Campione dei rispondenti e stima del totale popolazione con i pesi diretti**

Il secondo tabulato (cfr. *figura 5.21, capitolo precedente*) presenta per ogni variabile **Popolazioni pianificate utilizzate per lo stimatore** le seguenti statistiche:

47. **Pop. Pianif. stimatore**: indica le modalità assunte dalla variabile **Popolazioni pianificate utilizzate per lo stimatore**. L'ultima modalità della variabile **TOTALE** codifica la popolazione obiettivo. Tale variabile identifica le righe della tavola;
48. **nr. progr. popolaz.**: ricodifica, attraverso un numero progressivo, delle modalità della variabile **Popolazioni pianificate utiliz-**

**zate per lo stimatore.** L'ultima modalità della variabile codifica la popolazione obiettivo;

49. **Campione rispondenti (nr. unità rilevate):** indica la numerosità del campione rispondente, che presenta la stessa modalità della prima colonna. L'ultima riga, denominata **TOTALE**, è la somma delle righe precedenti;
50. **Somma dei pesi diretti:** indica la somma dei coefficienti iniziali, calcolata sulle unità del campione rispondente. Se i coefficienti sono stati corretti per tenere conto di eventuali unità non rispondenti, tali somme rappresentano stime non distorte delle corrispondenti sottopopolazioni individuate dalle modalità presenti nella prima colonna. L'ultima riga, denominata **TOTALE**, è la somma delle righe precedenti.

**Tab. 3: Controllo su Popolazioni pianificate che non hanno dati campionari**

Il terzo tabulato (cfr. *figura 5.22, capitolo precedente*) evidenzia i valori dei totali noti delle variabili ausiliarie per quelle sottopopolazioni pianificate che non hanno presentato alcuna unità campionaria. Il tabulato presenta un numero di colonne variabile e pari al numero di variabili ausiliarie utilizzate. Le informazioni presenti nel tabulato sono le seguenti:

51. **Pop. Pianif. stimatore:** indica la modalità della variabile **Popolazioni pianificate utilizzate per lo stimatore**, che identifica la sottopopolazione pianificata in cui non si è osservata alcuna unità rispondente;
52. **Totale\_noto\_variabale\_1:** indica il totale noto della prima variabile ausiliaria selezionata nella maschera di lancio della procedura, nella relativa sottopopolazione pianificata identificata dalla modalità della variabile **Pop. Pianif. stimatore**;
53. **Totale\_noto\_variabale\_i:** nella tabella sono presenti altre colonne in cui varia l'indice della variabile. La colonna i-esima indica il totale noto della i-esima variabile ausiliaria selezionata nella maschera di lancio della procedura, nella relativa sottopopolazione pianificata identificata dalla modalità della variabile **Pop. Pianif. stimatore**.

## 7. I file di output della funzione di Riponderazione di Genesees

*Il software produce alcuni data-set di output e alcuni file ascii, scritti sulla cartella di output scelta dall'utente, e produce infine il file "genesees.log".*

E' possibile memorizzare in file esterni le tabelle create dal software.

I file sono i seguenti:

- stampa1.txt,
- stampa2.txt,
- stampa3.txt,
- stampa4.txt,
- stampa5.txt,
- stampa6.txt,

Tali file vengono scritti **solo a richiesta** dell'utente quando - selezionata la stampa - egli utilizza l'opzione che permette la stampa in formato ascii (bottone "file", cfr. *figura 5.14*).

Ciascun file contiene una stampa, ad eccezione dei file stampa6.txt, che contiene tre tabulati: l'ultimo tabulato appare solo se il *data-set* dei "Totali Noti" si riferisce ad una o più popolazioni pianificate che non trovano riscontro nel *data-set* dei "Dati Campionari". Il software dunque crea il *tabulato 3* nel file stampa6.txt, che presenta le popolazioni pianificate che non hanno dati campionari (ovvero possibili mancate risposte) e scrive le corrispondenti informazioni nel *data-set* VUOTI (*cfr. capitolo 3, Sezione II*).

Per migliorare la leggibilità delle stampe è conveniente:

- aprire tali file con Microsoft Word
- selezionare tutto il testo e convertirlo in SAS Monospace, punto 8.

Il **file di log** contiene le informazioni che appaiono nella finestra di log del SAS ed è il seguente:

- `genesees.log`

Il SAS durante le elaborazioni, permette la visualizzazione delle informazioni di esecuzione sulla finestra di *Log*. L'esecuzione del software Genesees crea un *Log*, che - data la sua lunghezza e complessità - viene registrato su un file esterno, nella cartella di output, con il nome "**genesees.log**".

Ciò è particolarmente utile nel caso di un messaggio di errore: le informazioni memorizzate sono visualizzabili anche successivamente. Per leggere il file `genesees.log` è necessario terminare l'esecuzione della procedura e uscire dal software.

I data-set di output interessanti per l'utente sono i seguenti<sup>3</sup>:

**a) Data-set di lavoro:**

SAVESTIME creato per memorizzare parametri di input;  
NOTI\_MISS, CODICI\_DOPPI, CSENZAT, MISSING,  
VUOTI, creati per memorizzare gli errori rilevati sull'input.

**b) Data-set contenenti informazioni sui pesi campionari:**

PESIFIN, STAT, STIMEDIR, STIMEFIN.

Le informazioni contenute nei **data-set** sopra elencati verranno approfondite nel **capitolo 2** della **Sezione II**.

---

<sup>3</sup>

La cartella di output scelta dall'utente corrisponde alla libreria "outstime". Se, ad esempio, l'utente sceglie la cartella `c:\utente` - prendendo in considerazione il data-set di output PESIFIN - la procedura crea il data-set Sas di output "outstime.pesifin" che corrisponde al file `c:\utente\PESIFIN.sas7bdat` (data-set sas v.8) registrato nella cartella `c:\utente`. Per semplificare l'esposizione successiva si farà riferimento ai data-set solo con il nome, senza l'estensione del file o la libreria di riferimento.

# SEZIONE II

**Approfondimenti sulla costruzione  
dell'input e sui data-set di output  
della funzione di Riponderazione  
di Geneseees V. 3.0**



# 1. La costruzione del data-set di input della funzione di Riponderazione

***Sintesi:** Nel paragrafo 1.1 vengono descritte le variabili che devono essere contenute nei data-set di input (1.1.1) e i parametri richiesti per l'uso della funzione di riponderazione (1.1.2). Nel paragrafo 1.1.3 vengono presentati i controlli che il software effettua in automatico sui data-set di input per segnalare vincoli non rispettati nella costruzione delle variabili di input.*

*Nei paragrafi 1.2 e 1.3 si illustra la definizione di alcune variabili di input in relazione al tipo di stimatore utilizzato; in particolare, nel paragrafo 1.2.1, si illustra come identificare le variabili che specificano il gruppo di riferimento del modello.*

*Nel paragrafo 1.4 si pone l'attenzione alla determinazione dei valori di alcuni parametri richiamati nelle maschere di lancio del software, ovvero alla scelta della funzione di distanza.*

*Il paragrafo 1.5 approfondisce l'uso del software per calcolare i coefficienti finali di riporto in presenza di unità campionate non rispondenti.*

## 1.1 Le variabili ed i parametri di input

Il funzionamento del software prevede la definizione di alcune variabili di input e di alcuni parametri.

Le **variabili dei data-set di input** corrispondono a:

Variabili del data-set “Totali Noti”:

- Popolazioni pianificate utilizzate per lo stimatore
- Totali noti

Variabili del data-set “Dati Campionari”:

- Popolazioni pianificate utilizzate per lo stimatore
- Codice identificativo dell'unità campionaria



- Variabili ausiliarie
- Peso diretto
- Peso distanza

L'utente deve poi scegliere i seguenti **parametri**:

- una **funzione di distanza** e, ove necessario, i relativi coefficienti moltiplicatori dei limiti dell'intervallo di variazione del peso finale;
- una specifica popolazione pianificata utilizzata per lo stimatore, nel caso si vogliano calcolare i pesi finali con riferimento ad una sola popolazione.

Nel *paragrafo 1.1.1* vengono presentate le **variabili** del *data-set* di input, da definire per il calcolo dei pesi finali: il software richiede la presenza obbligatoria di alcune variabili di input e richiede specifici formati (in altre parole le variabili devono essere definite rigorosamente di tipo alfanumerico o di tipo numerico, come è di seguito indicato).

Nel *paragrafo 1.1.2* vengono descritti i **parametri** di input del software, specificandone il significato.

Infine nel *paragrafo 1.1.3* vengono segnalati i controlli effettuati dal software sulla costruzione delle variabili di input.

### **1.1.1 Le variabili di input**

**Attenzione! Il nome di tutte le variabili di input non può eccedere gli 8 caratteri!**

Per facilitare l'utente vengono di seguito presentate le variabili dei *data-set* di input, evidenziandone il significato in termini generali; viene inoltre presentata una **scheda riassuntiva delle caratteristiche richieste da ciascuna delle variabili** per il funzionamento del software. Infatti il software richiede la presenza obbligatoria di alcune variabili nel *data-set* di input e richiede specifici formati: le variabili devono cioè essere definite rigorosamente di tipo alfanumerico o di tipo numerico come viene di seguito indicato.

## Data-set dei totali noti

Nome colonna	Tipo	Lungh...	Formato	
TX1	Numero	8		← 2 Var. "Totali Noti": TX1
TX2	Numero	8		
DOMINIO	Testo	15		← Var. "Pop.pianif":

Il primo *data-set* contiene i valori dei totali noti con riferimento a ciascuna popolazione pianificata utilizzata per lo stimatore.

	TX1	TX2	DOMINIO
1	22849	866159	111
2	5860	212751	112
3	4661	167483	113
4	3434	662588	121
5	559	102172	122
6	151	77380	123

### 1) Popolazioni pianificate utilizzate per lo stimatore:

la variabile serve ad identificare una suddivisione della popolazione in diversi gruppi, rispetto alla quale sono noti i totali delle variabili ausiliarie utilizzate nello stimatore di calibrazione. Il *data-set* dei totali noti ha tante righe (ovvero tanti record) quanti sono i codici assunti da tale variabile.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: alfanumerico
Valori da assumere: qualsiasi
Numero di variabili: 1
Obbligatoria
Lunghezza: al massimo 15 caratteri.

### 2) Totali noti:

rappresentano i totali delle variabili ausiliarie. Il numero delle variabili "Totali noti" corrisponde al numero delle variabili ausiliarie utilizzate nello stimatore di calibrazione. Ciascuna variabile assume tanti valori quanti sono i codici assunti dalla variabile "Popolazioni pianificate utilizzate per lo stimatore".

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1 o più
1 almeno Obbligatoria

## Data-set dei dati campionari

Nome colonna	Tipo	Lungh...	Forma	
<sup>122</sup> <sub>67</sub> CK	Numero	8	←	Var. “Peso distanza”: CK
<sup>122</sup> <sub>67</sub> COEF	Numero	8	←	Var. “Peso diretto”: COEF
<sup>122</sup> <sub>67</sub> X1	Numero	8	} ←	2 Var. Ausiliarie X1 X2
<sup>122</sup> <sub>67</sub> X2	Numero	8		
<sup>Aa</sup> DOMINIO	Testo	15	←	Var. “Pop.pianif”:
<sup>Aa</sup> CODICE	Testo	15	←	Var. “Codice Un.”:

Il secondo *data-set* contiene informazioni riferite alle unità campionarie, ovvero ogni record si riferisce ad una diversa unità campionaria.

dominio	coef	x2	ck	x1	cod
111	35.925815532	31	1	1	1885
111	35.925815532	78	1	1	1886
111	35.925815532	29	1	1	1887
111	35.925815532	22	1	1	1895
111	35.925815532	20	1	1	1897
111	35.925815532	36	1	1	1904
111	35.925815532	43	1	1	1906
111	35.925815532	35	1	1	1907
111	35.925815532	121	1	1	1912
112	41.567896389	40	1	1	14
112	41.567896389	50	1	1	16
112	41.567896389	27	1	1	18
112	41.567896389	83	1	1	24

- 3) Codice dell'unità campionaria:  
rappresenta il codice identificativo delle unità campionarie.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1
Obbligatoria
Lunghezza: il numero può essere composto al massimo da 15 caratteri.

4) Il peso diretto:

la variabile indica il coefficiente diretto di riporto all'universo relativo all'unità elementare di campionamento. Nel caso si presentino mancate risposte totali, il peso diretto deve essere stato corretto precedentemente per tenerne conto.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1
Obbligatoria

5) Peso distanza:

è un peso da attribuire alla unità elementare di campionamento, ed è utile per definire lo specifico stimatore adottato; per approfondimenti si vedano i paragrafi successivi.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1
NON Obbligatoria

6) Variabili ausiliarie: queste variabili si riferiscono alle variabili utilizzate nello stimatore di calibrazione, di cui si conoscono i totali; nel *data-set* dei dati campionari contengono i valori assunti da ciascuna unità. Per approfondimenti sulla definizione di tali variabili nel *data-set* di input si vedano i paragrafi successivi.

Caratteristiche delle variabili del <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili di interesse: 1 o più
Obbligatorie

7) Popolazioni pianificate utilizzate per lo stimatore:

nel *data-set* dei dati campionari questa variabile indica a quale popolazione pianificata utilizzata per lo stimatore appartiene l'unità campionaria.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1
Obbligatoria

### 1.1.2 I parametri di input e la convergenza della procedura

**Primo parametro:** la selezione della popolazione pianificata

Nella selezione delle variabili di input per ciò che riguarda il data-set dei dati campionari, l'utente può scegliere di elaborare i pesi finali considerando solo le unità campionarie che appartengono ad una specifica popolazione pianificata utilizzata per lo stimatore (tramite la voce "Ulteriore Dominio", cfr. *paragrafo 5.2, Sezione I*).

Ciò è possibile solo se vi è stata una precedente elaborazione e il software è in grado di riconoscere i diversi codici relativi alla variabile "Popolazioni pianificate utilizzate per lo stimatore".

In tal caso – in automatico – il software mostra una lista dei codici e l'utente può selezionare la partizione di interesse. Ovviamente in questo caso il calcolo dei pesi finali è effettuato con riferimento alla sola partizione selezionata.

**Secondo parametro:** la selezione della funzione di distanza e la convergenza della procedura

La notazione simbolica che viene utilizzata nel seguito corrisponde a quella già utilizzata nel *capitolo 4, Sezione I*, dove sono stati introdotti alcuni concetti che vengono qui ripresi.

L'utente può scegliere una **funzione di distanza** tra quelle che mette a disposizione il software.

Genesees infatti dispone di sette possibili tipi di distanza per rendere possibile il raggiungimento dell'obiettivo, ovvero per identificare l'insieme dei pesi finali  $m_k$ , che modifichi il meno possibile - sotto determinati vincoli e sulla base appunto di una funzione di distanza prescelta - l'insieme dei

pesi diretti  $d_k$ . È da ricordare, infatti, che il software definisce gli stimatori nell'ambito della classe degli stimatori di ponderazione vincolata, i quali, a loro volta, si basano sulla scelta di una funzione di distanza (cfr. *capitolo 4, Sezione I*).

Per alcuni tipi di distanza, è necessario selezionare anche i **coefficienti moltiplicativi** dei limiti dell'intervallo di variazione del peso finale.

Per utilizzare le distanze **“Logaritmica troncata”** e **“Lineare troncata”** è necessario infatti definire i seguenti valori:

- 1) **coefficiente moltiplicativo del valore minimo di L** ( $0 = \text{coeff.molt.L} < 1$ )
- 2) **coefficiente moltiplicativo del valore massimo di U** (maggiore o uguale ad 1).

**Il software presenta i due valori di default pari a 0.5 e 1.5.**

Sull'uso delle distanze si consiglia di verificarne più di una, in quanto non tutte portano l'algoritmo iterativo alla convergenza.

Nel *capitolo 4* della *Sezione I* vengono evidenziate alcune considerazioni sull'uso di tali distanze. Per riportare qui le osservazioni di maggior rilievo, è da evidenziare che:

- a) La funzione di distanza euclidea può portare a pesi negativi o nulli, che generalmente non sono accettabili nella maggior parte delle applicazioni.  
E' però quella che richiede minor tempo di elaborazione, poiché non necessita di metodi iterativi per la soluzione.
- b) La funzione di distanza logaritmica porta a pesi sicuramente positivi che possono essere, tuttavia, estremamente alti ed in genere più elevati di quelli ottenuti con la distanza euclidea.
- c) La funzione logaritmica troncata è quella che ha il vantaggio di fornire pesi finali che assumono valori compresi in un intervallo che dipende dai valori dei coefficienti moltiplicativi (o moltiplicatori) sopra richiamati. Ciò perché esistono dei valori teorici da rispettare, ovvero i limiti inferiore e superiore del più piccolo intervallo teorico di variazione dei correttori intorno ad 1, per il quale il sistema di

minimo vincolato da risolvere ammette soluzione. Proprio su tale intervallo si agisce, variando i valori dei coefficienti moltiplicativi di cui sopra.

**Attenzione:** Nel caso in cui i coefficienti siano entrambi pari ad 1, l'intervallo effettivo coincide con l'intervallo minimo teorico e non si ammettono soluzioni.

Se si utilizza una distanza tra quelle per cui è possibile variare gli intervalli, è preferibile usare valori dei coefficienti vicini ad uno.

Nel caso in cui non si raggiunge la convergenza, è possibile provare ad allargare l'intervallo. In termini pratici, si inizia provando i valori di default; se poi i risultati non sono soddisfacenti, si prosegue variando l'intervallo, ad esempio considerandone uno molto ampio per poi restringerlo successivamente.

Una regola di tipo empirico può essere quella che consiste nell'attribuire al coefficiente moltiplicatore del valore minimo di  $L$  un valore iniziale prossimo allo zero e a quello di  $U$  un valore iniziale molto elevato (ed esempio 100); si procede poi, per tentativi successivi, aumentando l'uno e diminuendo l'altro verso l'unità.

Al termine dell'elaborazione è necessario verificare la **convergenza** della procedura iterativa, operazione possibile tramite l'analisi delle stampe di output. In particolare la *Tavola 1* (cfr. *figura 5.15, capitolo 5, capitolo 6, Sezione I*) presenta in caso di convergenza i valori di "Minima differenza stime" e "Massima differenza stime" pari a 0 e la *Tavola 5* (cfr. *figura 5.19, capitolo 5, Sezione I*) presenta le "Differenza stime finale" anche esse nulle.

Si osservi infine che tanto più i totali noti (*Tavola 5*) differiscono dalla corrispondenti stime dirette, tanto più i rapporti (*Tabulato 1*) sono diversi dall'unità, e quindi più forte sarà la correzione da apportare ai pesi diretti.

Le differenze in oggetto possono essere dovute a differenti cause, quali ad esempio, l'effetto delle mancate risposte totali o della sottocopertura della lista di selezione oppure della alta varianza campionaria delle stime dirette costruite su poche unità campionarie. In quest'ultimo caso ha senso ridefinire il sistema dei vincoli, aggregando tra loro quei totali noti le cui corrispondenti stime dirette sono basate su poche unità campionarie.

Nel caso particolare in cui più gruppi di totali noti si riferiscono, ciascuno indipendentemente dall'altro, alle stesse unità campionarie, un controllo a carico dell'utente consiste nel sommare i totali noti nei diversi gruppi e verificare che si raggiunga lo stesso totale di popolazione.

Infatti, se tali somme portassero a valori differenti, allora la convergenza non verrebbe mai raggiunta.

### **1.1.3 I controlli sui data-set di input**

Il software controlla e corregge automaticamente alcuni errori che l'utente ha originato nel costruire i *data-set* di input. In altri casi il software si blocca, segnalando l'errore.

#### **I controlli sul data-set dei totali noti**

- primo caso (correzione automatica):  
Nel caso di valori mancanti assunti da una delle variabili "Totali noti", tali valori sono sostituiti con un valore pari ad 0.
- secondo caso (fine dell'elaborazione):  
Nel caso di valori mancanti assunti dalla variabile "Popolazioni pianificate utilizzate per lo stimatore" nel *data-set* dei totali noti, il software si blocca e manda una segnalazione di errore.  
La variabile "Popolazioni pianificate utilizzate per lo stimatore" contenente i valori mancanti è scritta nel *data-set* NOTI\_MISS (cfr. capitolo 2).

#### **I controlli sul data-set dei dati campionari**

- primo caso (correzione automatica):
  - a) se il software trova un valore mancante assunto dalla variabile "Peso distanza", tale valore è sostituito con un valore pari ad 1;
  - b) se il software trova un valore mancante assunto dalla variabile "Peso diretto", tale valore è sostituito con un valore pari ad 1;
  - c) se il software trova un valore mancante assunto da una delle variabili ausiliarie, tale valore è sostituito con un valore pari ad 0.



- secondo caso (fine dell’elaborazione):

Il software si blocca e manda una segnalazione di errore se ci sono valori mancanti assunti dalla variabile “Codice”.

La variabile “Codice” contenente i valori mancanti è scritta nel data-set MISSING (cfr. capitolo 2).

Il software controlla che nel data-set dei dati campionari ciascuna unità campionaria sia identificata da un valore diverso della variabile “Codice”. Non possono perciò esistere record identificati da uno stesso valore della variabile “Codice”, che rappresenta una chiave univoca. In questo caso il software crea il data-set CODICI\_DOPPI.

## **Il controllo della corrispondenza tra i valori dei due data-set di input**

Come descritto nel *paragrafo 1.1*, in entrambi i *data-set* deve essere definita la variabile “Popolazioni pianificate utilizzate per lo stimatore”.

- Può accadere che per una certa popolazione pianificata di cui è noto il totale (e che dunque esiste nel *data-set* dei totali noti) non vi sia alcuna unità corrispondente nel *data-set* dei dati campionari. In questo caso il software non viene bloccato, ma procede con l’elaborazione. E’ infatti probabile che non si tratti di un errore, ma che nessuna delle unità selezionate nel campione appartenga a tale popolazione. Per segnalare tali casi, viene creato il *data-set* VUOTI (cfr. capitolo 2).
- Quando il software verifica che nel *data-set* dei dati campionari, una unità campionaria appartiene ad una popolazione pianificata che non esiste nel *data-set* dei totali noti, l’elaborazione viene bloccata. Per segnalare tali casi, viene creato il *data-set* CSENZAT (cfr. capitolo 2). (attenzione: tale *data-set* viene scritto anche nel caso in cui una popolazione assuma un valore mancante, in quanto il software non riconosce il corrispondente valore nel *data-set* dei totali noti).

## **1.2 Definizione delle variabili di input per un dato stimatore**

Il software calcola i coefficienti finali di riporto all’universo per gli stima-

tori appartenenti alla classe degli stimatori di *ponderazione vincolata* o *calibrazione*. A tale famiglia appartengono tutti i principali stimatori che utilizzano informazioni ausiliarie quali, ad esempio, gli stimatori *rapporto*, *rapporto post-stratificato*, *raking* e *regressione generalizzata* (per gli aspetti metodologici cfr. *appendice A.2*).

Per calcolare i coefficienti di uno specifico stimatore l'utente deve in primo luogo intervenire sulla definizione e la costruzione delle variabili del *data-set* “*totali noti*”, di seguito denominato TOTINP, e del *data set* “*dati campionari*”, di seguito denominato INP, introdotti nel *paragrafo 1.1*. Tali variabili sono indicate rispettivamente nelle *tabelle 1.1* e *1.2*.

**Tabella 1.1: Variabili del data set di input “totali noti” (TOTINP) necessarie per definire i coefficienti finali di riporto dello stimatore campionario che si intende adottare.**

Variabili di input (paragrafo 1.1)	Nomi sintetici delle variabili adottati nel testo
Popolazioni pianificate utilizzate per lo stimatore	POP_PIAN
Totali noti	TX1, ..., TXj, ..., TXJ

**Tabella 1.2: Variabili del data-set di input “dati campionari” (INP) necessarie per definire i coefficienti finali di riporto dello stimatore campionario che si intende adottare.**

Variabili di input (paragrafo 1.1)	Nomi sintetici delle variabile adottati nel testo
Identificativo unità campionaria	IDEN
Peso diretto	COEF_DIR
Variabili ausiliarie	X1, ..., Xj, ..., XJ
Popolazioni pianificate utilizzate per lo stimatore	POP_PIAN
Peso distanza	CK

Oltre alla definizione di tali variabili nei due *data-set*, per definire correttamente lo stimatore da utilizzare devono, infine, essere impostati alcuni parametri, richiamati attraverso le maschere di lancio della procedura informatica (cfr. capitolo 5, *Sezione I*).

Il *paragrafo 1.2.1* illustra gli aspetti relativi alla definizione e alla

costruzione delle variabili POP\_PIAN, TX1, ..., TXj, ..., TXJ (*tabella 1.1*), appartenenti al *data-set* TOTINP e delle variabili POP\_PIAN e X1, ..., Xj, ..., XJ (*tabella 1.2*), presenti nel *data-set* INP. Nella terminologia propria della teoria degli stimatori di ponderazione vincolata, con tali variabili si specificano i *gruppi di riferimento* dello stimatore (per approfondimenti cfr. *appendice A.2*).

I *paragrafi 1.2.1, 1.2.2, 1.2.3, 1.2.4 e 1.3* pongono l'attenzione ai valori che devono assumere le variabili ausiliarie X1, ..., Xj, ..., XJ, TX1, ..., TXj, ..., TXJ, e le variabili IDEN e CK. Nel *paragrafo 1.4* si pone l'attenzione alla determinazione dei valori di alcuni parametri, richiamati nelle maschere di lancio del software, per ottenere i coefficienti finali di alcuni degli stimatori più noti in letteratura.

Il *paragrafo 1.5* approfondisce l'uso del software per calcolare i coefficienti finali di riporto in presenza di unità campionate non rispondenti.

L'utente interessato al calcolo dei coefficienti finali di riporto degli stimatori *rapporto*, *rapporto post-stratificato*, *ratio raking* può consultare direttamente il *paragrafo 1.3*. Per quanto riguarda gli stimatori *raking generalizzato*, di *regressione generalizzata* o gli stimatori più generali di *ponderazione vincolata* o *calibrazione* (per la definizione di tali stimatori si può vedere l'*appendice A.1*) è consigliata anche la lettura dei *paragrafi 1.2.1 e 1.2.2 e 1.2.3*.

È utile, infine, sottolineare che i due *data-set* di input TOTINP e INP non presentano esplicitamente informazioni sul disegno di campionamento adottato dall'utente. Tali informazioni sono contenute nel coefficiente iniziale COEF\_DIR del data set INP. È, dunque, essenziale che i coefficienti di input siano calcolati correttamente in quanto il software non effettua alcun controllo.

### **1.2.1 Gruppo di riferimento del modello**

Gli stimatori di ponderazione vincolata si fondano sul concetto di *gruppo di riferimento*<sup>4</sup> (per approfondimenti cfr. *appendice A.1 e A.2*) il quale, in sin-

---

<sup>4</sup> L'espressione "gruppo di riferimento del modello" ha origine dalla terminologia adottata per lo stimatore di regressione generalizzata in cui il gruppo di riferimento è una sottogruppo del campione in cui si stima il modello di regressione

tesi, è una sottopopolazione per la quale sono noti i totali di alcune variabili ausiliarie<sup>5</sup> utilizzate nella strategia campionaria per cercare di migliorare l'efficienza delle stime e compensare gli errori di copertura e di mancata risposta totale. Il processo di calibrazione tiene conto di questi totali generando dei coefficienti finali che se applicati per la stima del totale di una tra le variabili ausiliarie che hanno originato gli stessi coefficienti finali riproducono esattamente il totale noto per il dato gruppo di riferimento.

Per indicare al software quali sono i gruppi di riferimento che si intende utilizzare è necessario strutturare i *data-set* di input nel modo opportuno. Al fine di descrivere la costruzione dei *data-set* è necessario formulare una premessa.

Data la popolazione dalla quale si estrae il campione si può definire una sottopopolazione in relazione alla stratificazione del disegno. In tal caso si possono individuare due tipi di sottopopolazioni: le *sottopopolazioni pianificate* e le *sottopopolazioni non pianificate*.

Le sottopopolazioni pianificate, definibili quando si adotta un disegno stratificato, sono costruite in modo tale da coincidere con uno o più strati del disegno. In tal caso, dato uno strato o un insieme di strati, tutte le unità dello strato o dell'insieme di strati appartengono ad una ed una sola sottopopolazione pianificata.

Le sottopopolazioni non pianificate, invece, sono costruite in modo tale che le unità di un generico strato appartengono solo in parte ad una generica sottopopolazione. In generale, se la sottopopolazione non pianificata è costituita da unità provenienti da strati diversi, è importante che per almeno uno strato non siano presenti tutte le unità nella sottopopolazione perché questa possa definirsi non pianificata.

I gruppi di riferimento, essendo sottopopolazioni in cui sono noti dei totali per alcune variabili ausiliarie, si possono classificare come pianificati e non pianificati. In particolare, procedendo ad una descrizione più dettagliata si hanno le tre categorie seguenti:

---

<sup>5</sup> Il software ammette solo variabili ausiliarie quantitative o qualitative di tipo dicotomico. Gli altri tipi di variabili qualitative devono essere trasformate in forma dicotomica. I dettagli per questa trasformazione sono illustrati nel paragrafo A.3.2 dell'appendice A.3

- (i) sottopopolazioni pianificate;
- (ii) sottopopolazioni non pianificate definite all'interno di strati o aggregazioni di strati;
- (iii) sottopopolazioni non pianificate;

Per quanto riguarda il caso (i), si ha che ciascun gruppo di riferimento può essere formato da:

- (A1) tutte le unità appartenenti ad un singolo strato del disegno;
- (A2) tutte le unità appartenenti ad un'aggregazione di strati del disegno;
- (A3) l'intera popolazione di riferimento (in questo caso si ha un'unica sottopopolazione pianificata che coincide con la popolazione stessa).

Considerando invece il caso (ii), ciascun gruppo di riferimento può contenere:

- (B1) una parte delle unità contenute in uno strato;
- (B2) una parte delle unità contenute in una aggregazione di strati;

Infine per il caso (iii), ciascun gruppo di riferimento deve essere composto da:

- (B3) una parte delle unità contenute nella popolazione (che non coincide con l'insieme completo di unità contenute in uno strato o in una aggregazione di strati).

I  $D$  gruppi di riferimento, che costituiscono una partizione della popolazione, formata secondo uno dei sei criteri sopra illustrati, risultano allora pari a:

- (A1)  $D=H$ , in cui  $H$  rappresenta il numero complessivo degli strati;
- (A2)  $D=H_G (<H)$ , in cui  $H_G$  è il numero degli insiemi di strati aggregati;
- (A3)  $D=1$ ;
- (B1)  $D=H \times Q$ , in cui  $Q$  è il numero delle sottopopolazioni non pianificate presenti all'interno dello strato. Queste sottopopolazioni devono essere definite allo stesso modo in ciascuno strato;
- (B2)  $D=H_G \times Q$ , in cui  $Q$  è il numero delle sottopopolazioni non pianificate presenti all'interno di una aggregazione di strati. Queste sottopopolazioni devono essere definite allo stesso modo in ciascuna aggregazione di strati;

(B3)  $D=Q$ , in cui  $Q$  è il numero delle sottopopolazioni non pianificate presenti all'interno della popolazione di studio.

Definendo i gruppi di riferimento secondo uno dei sei punti precedenti, si può osservare che le partizioni ottenute con i punti (A1), (A2) e (A3) rappresentano casi particolari rispettivamente dei punti (B1), (B2) e (B3) quando si ha  $Q=1$ . Quando si costruiscono i *data-set* di input l'utente deve tenere in considerazione queste tipologie di gruppi di riferimento.

### 1.2.2 Stimatore definito sui totali noti di una sola variabile ausiliaria

Gli stimatori di ponderazione vincolata considerati nel paragrafo tengono conto dei totali di una variabile ausiliaria  $x$  (per approfondimenti sul tipo di variabile ausiliaria cfr. *appendice A.3*) noti su un insieme di  $D$  gruppi di riferimento definiti combinando una variabile di stratificazione  $s$  (con modalità  $b=1, \dots, b, \dots, H$ )<sup>6</sup> e una variabile  $v$  (con modalità  $q=1, \dots, q, \dots, Q$ ) che non contribuisce alla stratificazione del disegno. I  $D (=H \times Q)$ <sup>7</sup> gruppi costituiscono una partizione  $P$  della popolazione. Gli stimatori che presentano una tale struttura di totali noti sono quello del *rapporto separato* (in cui  $Q=1$ ), del *rapporto post-stratificato* (in cui  $H=1$ ), del *rapporto post-stratificato separato* e dello stimatore del *rapporto post-stratificato combinato* (in cui  $H=1$ ). Rientrano in questa classe, considerando il caso  $D=1$ , anche lo stimatore di *Hajek* (stimatore del rapporto che utilizza la numerosità della popolazione come totale noto), del *rapporto semplice*, *rapporto combinato* e di *regressione semplice*.

Per indicare al software quali sono i gruppi di riferimento è necessario definire correttamente alcune variabili del *data-set* TOTINP e INP (cfr. *tabelle 1.1 e 1.2*).

In particolare per il primo *data-set* è necessario fare attenzione alla definizione delle variabili POP\_PIAN e TX1, ..., TXJ e per il secondo *data-set* bisogna considerare le variabili POP\_PIAN e le variabili, X1, ..., XJ.

---

<sup>6</sup> Per brevità non si descrive il caso in cui la variabile  $s$  è costituita da modalità che sono aggregazioni degli strati del disegno. Questo caso è facilmente ricavabile dalle considerazioni sviluppate nel paragrafo e viene ripreso nell'appendice A.3

<sup>7</sup> In generale i gruppi di riferimento si possono identificare anche combinando una variabile che rappresenta una aggregazione degli strati del disegno (si veda A.3)

Una tale struttura di totali noti si presenta nell'esempio seguente.

*Esempio 1.1:*

*Si consideri un campione di individui, da una popolazione di 1.050 unità, stratificato secondo la variabile sesso (variabile  $s$ ) e sia noto il totale degli individui per ogni combinazione delle variabili sesso e classe di età (variabile  $v$ ). In questo caso la variabile ausiliaria  $x$  è una variabile sempre pari ad 1.*

*Nella tabella 1.3 sono descritte le due variabili.*

**Tabella 1.3 – Descrizione delle variabili che definiscono lo stimatore nell'esempio 1.1**

Variabile	Modalità della variabile	Numero modalità	Simbolo	Numero delle modalità (simbolo)
Sesso	Uomo; Donna.	2	S	S
Classe di età	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre.	4	V	Q

*Nella tabella 1.4, sono presentati i rispettivi totali noti per ciascuno dei ( $D = Q \times H$ ) gruppi di riferimento.*

**Tabella 1.4 – Totali noti dello stimatore di ponderazione vincolata nell'esempio 1.1**

Classe di età Sesso	0-14 anni	15-34 anni	35-54 anni	55 anni e oltre
Uomo	50	200	200	30
Donna	40	140	270	120

*Per considerare correttamente questa struttura di totali noti il data-set TOTINP può essere costruito secondo diverse alternative.*

*La prima alternativa, denominata schema A, prevede che in TOTINP sia presente un solo record e ci siano  $J (= D)$  variabili  $TX1, \dots, TXJ$ .*

*Nell'esempio si definisce, pertanto, una variabile che individua la variabile POP\_PLAN (nell'esempio è stata chiamata A) che assume una modalità arbitraria, mentre le variabili  $TX1, \dots, TX8$  rappresentano gli otto gruppi di riferimento e presentano come valori i totali noti. Ad esempio (cfr. figura 1.1) la variabile  $TX1$  presenta il totale noto del gruppo di riferimento individuato dalla combinazione delle modalità uomo e 0-14 anni.*

**Figura 1.1 – Costruzione del data-set TOTINP (totali noti) secondo lo schema A**

VIEWTABLE: Work.Esempiotot									
	A	TX1	TX2	TX3	TX4	TX5	TX6	TX7	TX8
1	cost	50	200	200	30	40	140	270	120

In base a questa costruzione di TOTINP il data-set INP deve essere costruito come è illustrato nella figura 1.2.

**Figura 1.2 – Costruzione del data-set INP secondo lo schema A**

VIEWTABLE: Work.Esempioinp												
	cod	A	X1	X2	X3	X4	X5	X6	X7	X8	...	
1	1	cost	0	0	0	0	0	1	0	0	...	
2	2	cost	0	0	0	1	0	0	0	0	...	
3	3	cost	0	0	0	0	0	0	0	1	...	
4	4	cost	1	0	0	0	0	0	0	0	...	
5	5	...	1	0	0	0	0	0	0	0	...	

Nel data-set per ciascun record, identificato dalle modalità della variabile COD (IDEN secondo la tabella 1.2), sono presenti tra le altre la variabile A e le variabili X1, ..., X8. Per quanto riguarda la variabile A essa presenta un'unica modalità su tutti i record, uguale a quella imposta nel data-set TOTINP. Per quanto riguarda le variabili ausiliarie, deve esistere una corrispondenza tra le coppie TX1 e X1, ..., TX8 e X8.

Ad esempio la variabile X1 individua il gruppo di riferimento descritto dalla combinazione delle modalità uomo e 0-14 anni, in accordo con quanto definito sul data-set TOTINP.

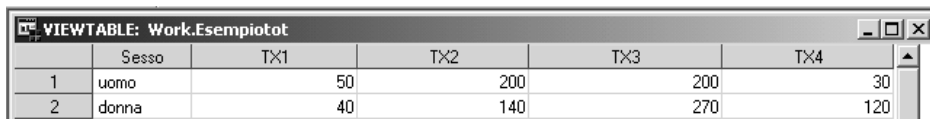
Inoltre, il criterio per assegnare i valori a tali variabili è il seguente: se il record appartiene al gruppo di riferimento individuato dalla generica variabile Xj tale variabile assume il valore osservato della variabile x (in questo caso il valore è 1); se il record non appartiene al gruppo di riferimento individuato da Xj tale variabile assume valore nullo. Ad esempio, il primo record della figura 1.2, è una donna con età 15-34 anni.

La seconda alternativa, denominata schema B, prevede che nel data-set TOTINP ciascun gruppo di riferimento sia individuato dalla combinazione della modalità di riga



della variabile POP\_PLAN e dalla variabile ausiliaria presente in colonna. Nell'esempio riportato nella figura 1.3., la cella è identificata dalla combinazione della modalità di riga della variabile sesso, che assume il ruolo della variabile POP\_PLAN, con una delle colonne identificate dalle variabili TX1, ..., TXJ, che sono pari al numero delle modalità della variabile classe di età ( $J = Q$ ).

**Figura 1.3 – Costruzione del data-set TOTINP (totali noti) secondo lo schema B**

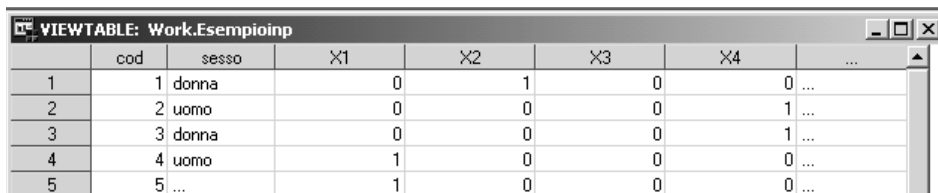


	Sesso	TX1	TX2	TX3	TX4
1	uomo	50	200	200	30
2	donna	40	140	270	120

Nella figura la cella individuata dalla riga uomo e la variabile TX1, individua il gruppo di riferimento uomo 0-14 anni. Ad ogni variabile TXj si associa, quindi, una modalità della variabile classe di età.

Secondo lo schema B, il data-set INP, nell'esempio proposto presenta le caratteristiche illustrate nella figura 1.4.

**Figura 1.4 – Costruzione del data-set INP secondo lo schema B**



	cod	sesso	X1	X2	X3	X4	...
1	1	donna	0	1	0	0	...
2	2	uomo	0	0	0	1	...
3	3	donna	0	0	0	1	...
4	4	uomo	1	0	0	0	...
5	5	...	1	0	0	0	...

Secondo questo schema la variabile sesso assume la modalità osservata sul record corrispondente.

Per quanto riguarda le variabili ausiliarie, deve esistere una corrispondenza tra le coppie TX1 e X1, ..., TX4 e X4.

Ad esempio la variabile X1 individua in questo caso la modalità 0-14 anni, in accordo con quanto definito sul data-set TOTINP.

Inoltre il criterio per assegnare i valori a tali variabili è il seguente: se il record presenta la classe di età individuata dalla generica variabile Xj, tale variabile assume il valore

osservato della variabile  $x$  (in questo caso il valore è 1); altrimenti la generica variabile  $X_j$  assume valore nullo. Ad esempio, il primo record della figura 1.4., è una donna con età 15-34 anni.

Alcune importanti indicazioni che si possono trarre da questo esempio, e che sono sempre valide nelle costruzioni dei data-set, sono le seguenti:

- la variabile POP\_PIAN può assumere come modalità solo quelle che definiscono gli strati (come nello schema B dell'esempio 1.1) o modalità di variabili che rappresentano aggregazione di strati (in particolare lo schema A rappresenta un caso estremo di aggregazione di tutti gli strati del disegno). Ciascuna modalità di POP\_PIAN identifica, pertanto, una sottopopolazione pianificata;
- le variabili TX1, ..., TXJ e X1, ..., XJ, individuano le modalità che combinate con quelle della variabile POP\_PIAN definiscono i gruppi di riferimento dello stimatore di ponderazione vincolata. Tali variabili da una parte devono essere definite attraverso le modalità delle variabili che definiscono i gruppi di riferimento ma non rientrano nella definizione della stratificazione; dall'altra possono essere definite considerando anche le modalità delle variabili che contribuiscono a definire la stratificazione del disegno (è questo il caso dello schema A presentato nell'esempio 1.1). Ciascuna variabile TXj e Xj può identificare, pertanto, una qualsiasi sottopopolazione, pianificata o non pianificata;
- è fondamentale fare attenzione all'ordine di inserimento delle variabili al momento del lancio della procedura, in quanto il software associa la prima variabile selezionata tra i totali noti, TXj, con la prima variabile selezionata del tipo Xj, ed analogamente accoppia la seconda, la terza (e così via) variabile TXj selezionata con la seconda, la terza (e così via) variabile Xj selezionata<sup>8</sup>.

Le regole generali per la costruzione dei *data-set* TOTINP e INP (rispettate nell'esempio 1.1) secondo lo schema A sono descritte nell'elenco seguente e nelle tabelle 1.5 e 1.6.

---

<sup>8</sup> Si ricorda che i nomi TXj e Xj assegnati alle variabili per definire i gruppi di riferimento sono utilizzati con un puro scopo descrittivo. In realtà non esistono vincoli particolari sul tipo di nome da assegnare, e quindi, a maggior ragione, non è necessario attribuire ai nomi delle variabili Totali noti e Variabili ausiliarie un indice che le possa associare a coppie

Caratteristiche del *data-set* TOTINP secondo lo schema A;

- un solo record nel *data-set*;
- la variabile POP\_PIAN con un solo valore scelto arbitrariamente dall'utente;
- un numero di variabili TXj corrispondente al numero dei gruppi di riferimento ( $j=1, \dots, d, \dots, D$ ). Ogni variabile TXj identifica uno specifico gruppo di riferimento. I valori assunti da tali variabili sono i rispettivi totali noti della variabile ausiliaria  $x$  per il corrispondente gruppo di riferimento.

Caratteristiche del *data-set* INP secondo lo schema A;

- la variabile POP\_PIAN risulta costante per ciascun record del *data-set* e pari al valore scelto in TOTINP;
- il numero delle variabili Xj corrisponde ai  $D$  gruppi di riferimento. Ogni variabile Xj ( $j=1, \dots, d, \dots, D$ ) identifica uno specifico gruppo di riferimento. Per ciascun record solo una di queste variabili Xj assume il valore della variabile  $x$  osservato sul record stesso, mentre le altre sono nulle. La variabile che presenta il valore di  $x$  è quella che identifica il gruppo di riferimento a cui appartiene il record stesso.

**Tabella 1.5 - Descrizione dello Schema A: definizione del data-set TOTINP con una variabile ausiliaria  $x$  e una partizione  $P$  con  $D$  gruppi di riferimento.**

POP_PIAN	TX1		TXj (con $j=d$ )		TXJ (con $J=D$ )
Costante	Tot. X sul primo gruppo di riferimento		Tot. X sul j-esimo gruppo di riferimento		Tot. X sul J-esimo (=D-esimo) gruppo di riferimento

**Tabella 1.6 - Descrizione dello Schema A: definizione del data-set INP con  $n$  record ( $i=1, \dots, n$ ) con una variabile ausiliaria  $x$  e una partizione  $P$  con  $D$  gruppi di riferimento. Esempio per il record  $i$  appartenente al  $j$ -esimo gruppo di riferimento.**

IDEN	POP_PIAN	X1		Xj( $j=d$ )		XJ( $J=D$ )	COEF_DIR	CK
1	...	...		...		...	...	...
...	...	...		...		...	...	...
$i$	Costante	0		$x$		0	Coeff. Diretto unità	Vedere par. 1.3 (Sez. II)
...	...	...		...		...	...	...
$n$	...	...		...		...	...	...

Le regole per la costruzione dei *data-set* TOTINP e INP secondo lo schema B sono presentate nell'elenco seguente e nelle tabelle 1.7 e 1.8.

Caratteristiche del *data-set* TOTINP secondo lo schema B;

- esistono  $S$  record nel *data-set* (con  $S=H$  oppure  $S=H_G$ ). Ciascun record è identificato da una modalità della variabile POP\_PIAN (ogni modalità si presenta una sola volta) ed individua una sottopopolazione pianificata (uno strato o una aggregazione di strati) ;
- sono presenti tante variabili TXj pari al numero delle modalità della variabile che definisce i gruppi di riferimento ma non rientra nella definizione della stratificazione del disegno. Per ciascun record del *data-set*, la variabile TXj assume il valore del totale noto della variabile ausiliaria  $x$  per il relativo gruppo di riferimento identificato dall'incrocio della modalità POP\_PIAN che presenta il record e la modalità individuata dalla variabile TXj.

Caratteristiche del *data-set* INP secondo lo schema B;

- la variabile POP\_PIAN presenta per ciascun record la modalità della variabile che definisce una sottopopolazione pianificata (uno strato o un'aggregazione di strati), in accordo con quanto definito nel *data-set* TOTINP, a cui appartiene il record stesso;
- sono presenti una serie di variabili Xj, ciascuna delle quali coincide con una modalità della variabile che definisce i gruppi di riferimento ma non rientra nella definizione della stratificazione del disegno. In particolare deve esistere una corrispondenza tra ciascuna coppia TXj e Xj. Considerando una generica Xj, questa assume il valore della variabile  $x$  osservata sul record stesso se il record appartiene al gruppo di riferimento identificato dalla combinazione della modalità assunta dalla variabile POP\_PIAN e da quella individuata dalla stessa variabili Xj, altrimenti la variabile Xj assume valore nullo.

**Tabella 1.7 - Descrizione dello Schema B: definizione del data-set TOTINP con una variabile ausiliaria  $x$  e una partizione  $P$  con  $D (=H \times Q)$  gruppi di riferimento.**

POP_PIAN	TX1	...	TXj (j=q)	...	TXJ(J=Q)
1	Tot. X sul primo gruppo di riferimento	...	...	...	...
...	...	...	...	...	...
$h$	...	...	Tot. X sul gruppo di riferimento definito dalla combinazione $h$ e $q$	...	...
...	...	...	...	...	...
$H$	...	...	...	...	Tot. X sul $D$ -esimo gruppo di riferimento

**Tabella 1.8 - Descrizione dello Schema B: definizione del data-set INP con  $n$  record ( $i=1, \dots, n$ ), una variabile ausiliaria  $x$  e una partizione  $P$  con  $D$  gruppi di riferimento. Esempio per il record  $i$  appartenente al  $d$ -esimo ( $d^\circ(h;q)$ ) gruppo di riferimento.**

IDEN	POP_PIAN	X1	...	Xj(j=q)	...	XJ(J=Q)	COEF_DIR	CK
1	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
$i$	$h$	0	...	$x$	...	0	Coef. Diretto unità	Vedere par. 1.3 (Sez. II)
...	...	...	...	...	...	...	...	...
$n$	...	...	...	...	...	...	...	...

In alcune occasioni si rende disponibile una terza alternativa, denominata schema C, per costruire i gruppi di riferimento. Ciò avviene quando i gruppi sono definiti con una variabile che definisce le sottopopolazioni pianificate  $s$  che è il risultato della combinazione di due o più variabili  $s_1, \dots, s_R$ . Di seguito è presentato un esempio con  $R=2$ .

*Esempio 1.2:*

*Si consideri un campione di individui, da una popolazione di 1.050 unità, stratificato secondo la variabile combinata (variabile  $s$ ) sesso (variabile  $s_1$ ) e ripartizione*

geografica di residenza dell'unità campionaria (variabile  $s_2$ ) e sia noto il totale degli individui per ogni combinazione della variabile  $s$  e la variabile classe di età (variabile  $v$ ). In questo caso la variabile ausiliaria  $x$  è una variabile sempre pari ad 1. Nella tabella 1.9 sono descritte le due variabili, mentre nella tabella 1.10, sono presentati i rispettivi totali noti per ciascuno dei 24 ( $D = Q \times H$ ) gruppi di riferimento.

**Tabella 1.9 – Descrizione delle variabili che definiscono lo stimatore nell'esempio 1.2**

Variabile	Modalità della variabile	Numero modalità	Simbolo	Numero delle modalità (simbolo)
Sesso × Ripartizione geografica	Uomo×Nord; Uomo×Centro; Uomo×Sud; Donna×Nord; Donna×Centro; Donna×Sud;	6	$s$	$S$
Classe di età	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre.	4	$v$	$Q$

**Tabella 1.10 – Descrizione delle variabili che definiscono lo stimatore nell'esempio 1.2**

Classe di età Sesso	0-14 anni	15-34 anni	35-54 anni	55 anni e oltre
Uomo×Nord	30	80	100	15
Uomo×Centro	10	20	30	10
Uomo×Sud	10	100	70	5
Donna×Nord	20	60	120	50
Donna×Centro	10	60	100	40
Donna×Sud	10	20	50	30

Per descrivere come poter costruire l'input per il software, consideriamo per brevità il data-set TOTINP. Secondo lo schema A di devono definire 24 variabili  $TX_j$  mentre applicando lo schema B si hanno 4 variabili  $TX_j$ . Tuttavia avendo a disposizione una variabile di stratificazione combinata con due variabili che rientrano nella definizione dei gruppi di riferimento si può procedere ad una terza alternativa (schema C) per la costruzione dei data-set di input.

Si sceglie una tra la variabile sesso e ripartizione geografica. La variabile scelta, ad esempio la ripartizione geografica assume il ruolo della variabile POP\_PLAN. Si costruisce, quindi, una nuova variabile combinando la variabile non scelta precedentemente, nell'esempio la variabile sesso, con la variabile classe di età.

In tale modo si ottiene il data-set TOTINP illustrato nella figura 1.5., in cui le variabili  $TX_1, \dots, TX_4$  individuano le quattro classi di età per gli uomini, mentre le variabili  $TX_5, \dots, TX_8$ , sono relative alle quattro classi di età per le donne.

**Figura 1.5 – Costruzione del data-set TOTINP (totali noti) secondo lo schema C**

VIEWTABLE: Work.Esempiotot									
	ripartizione_geografica	TX1	TX2	TX3	TX4	TX5	TX6	TX7	TX8
1	nord	30	80	100	15	20	60	120	50
2	centro	10	20	30	10	10	60	100	40
3	sud	10	100	70	5	10	20	50	30

La figura 1.6., mostra come può presentarsi il data-set INP. Nella figura, ad esempio, il primo record è relativo ad una donna con età 15-34 anni residente nel sud.

**Figura 1.6 - Costruzione del data-set INP secondo lo schema C**

VIEWTABLE: Work.Esempioinp										
	cod	ripartizione_geografica	X1	X2	X3	X4	X5	X6	X7	X8
1	1	sud	0	0	0	0	0	1	0	0 ...
2	2	sud	0	0	0	1	0	0	0	0 ...
3	3	nord	0	0	0	0	0	0	0	1 ...
4	4	centro	1	0	0	0	0	0	0	0 ...
5	5	...	1	0	0	0	0	0	0	0 ...

In termini generali, quando la variabile che definisce le sottopopolazioni pianificate alla base dei gruppi di riferimento è composta da un insieme di due o più variabili, per definire i due *data-set*, con lo schema C, è necessario suddividere preventivamente queste variabili in due classi. Attraverso la combinazione delle modalità delle variabili appartenenti alla prima classe si definiscono le  $S_1$  modalità della variabile POP\_PIAN. Attraverso la combinazione delle  $S_2$  modalità della seconda classe di variabili e le  $Q$  modalità di una eventuale variabile che non rientra nella definizione della stratificazione del disegno si determinano le  $Q \times S_2$  variabili TX<sub>j</sub> e X<sub>j</sub>.

Come per lo schema B, la variabile che coincide con POP\_PIAN deve quindi rappresentare una combinazione di modalità delle variabili che contribuiscono alla stratificazione oppure la combinazione di aggregazione di tali modalità. In ogni caso viene rispettato il principio per cui ogni modalità della variabile POP\_PIAN identifica una sottopopolazione pianificata.

Nelle *tabelle 1.11 e 1.12*, sono illustrate sinteticamente le regole per la costruzione dei due *data-set* secondo lo schema C.

**Tabella 1.11 - Descrizione dello Schema C: definizione del data-set TOTINP con una variabile ausiliaria  $x$  e una partizione  $P$  con  $D$  gruppi di riferimento. Stratificazione del disegno secondo le modalità di una variabile  $s$  ottenuta come combinazione delle variabili  $s_1$  (con modalità  $h_1=1, \dots, H_1$ ) e  $s_2$  (con modalità  $h_2=1, \dots, H_2$ )**

POP_PIAN	TX1	...	TXj ( $j=h_2; q$ )	...	TXJ ( $J=H_2; Q$ )
1	Tot. $X$ sul primo gruppo di riferimento	...	...	...	...
...	...	...	...	...	...
$H_1$	...	...	Tot. $x$ sul gruppo di riferimento definito dalla combinazione $h_1$ e ( $h_2; q$ )	...	...
...	...	...	...	...	...
$H_1$	...	...	...	...	Tot. $X$ sul $D$ -esimo gruppo di riferimento

**Tabella 1.12 - Descrizione dello Schema C: definizione del data-set INP con  $n$  record ( $i=1, \dots, n$ ), una variabile ausiliaria  $x$  e una partizione  $P$  con  $D$  gruppi di riferimento. Stratificazione del disegno secondo le modalità di una variabile  $s$  ottenuta come combinazione delle variabili  $s_1$  (con modalità  $h_1=1, \dots, H_1$ ) e  $s_2$  (con modalità  $h_2=1, \dots, H_2$ ). Esempio per il record  $i$  appartenente al  $d$ -esimo ( $d^o(h_1; h_2; q)$ ) gruppo di riferimento.**

IDEN	POP_PIAN	X1	...	Xj ( $j=h_2; q$ )	...	XJ ( $J=H_2; Q$ )	COEF_DIR	CK
1	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
$i$	$h_1$	0	...	$x$	...	0	Coeff. Diretto unità	Vedere par. 1.3 (Sez. II)
...	...	...	...	...	...	...	...	...
$n$	...	...	...	...	...	...	...	...



### 1.2.3 Stimatore definito su diversi sistemi di totali noti

Il processo di calibrazione può ricorrere ai totali noti di più d'una variabile ausiliaria e per più partizioni dell'universo in gruppi di riferimento. Per alcuni stimatori, ad esempio, si possono avere a disposizione i totali di una variabile ausiliaria per diverse partizioni e partizioni sulle quali sono noti i totali di diverse variabili ausiliarie. Stimatori, noti in letteratura, che presentano questa struttura più generale dei totali noti sono il *ratio raking* (una variabile ausiliaria e due partizioni in gruppi di riferimento), il *raking generalizzato* e *regressione generalizzata* o, ancora più in generale, gli stimatori di *ponderazione vincolata*.

Assumendo, pertanto, di avere a disposizione  $x_1, \dots, x_t, \dots, x_T$ , variabili ausiliarie per le quali sono noti i totali su varie partizioni in gruppi di riferimento della popolazione obiettivo, la costruzione dei due *data-set* di input può seguire uno dei tre diversi schemi introdotti nel paragrafo precedente.

Caratteristiche del *data-set* TOTINP secondo lo schema A;

- esiste un solo record nel *data-set*;
- la variabile POP\_PIAN presenta un solo valore scelto arbitrariamente dall'utente;
- per ogni variabile  $x_t$  e per la partizione associata, in cui sono noti i totali si crea un insieme di variabili TXj. Il numero delle variabili TXj nell'insieme è dato dal numero di gruppi di riferimento che identificano la partizione in questione. La generica variabile TXj identifica, pertanto la variabile ausiliaria  $x_t$  ed un preciso gruppo di riferimento sul quale è noto il totale di  $x_t$ . Il valore che assume ciascuna variabile dell'insieme è il totale noto della variabile  $x_t$  nel gruppo di riferimento;
- si forma un insieme di variabili TXj per ogni coppia partizione – variabile  $x_t$  in cui sono noti i totali;

Caratteristiche del *data-set* INP secondo lo schema A;

- la variabile POP\_PIAN risulta costante per ciascun record del *data-set* e pari al valore scelto in TOTINP;
- per ogni variabile  $x_t$  e per la partizione associata, in cui sono noti i

totali si crea un insieme di variabili  $X_j$ . Il numero delle variabili  $X_j$  nell'insieme è dato dal numero di gruppi di riferimento che identificano la partizione. Per ciascun record solo una di queste variabili  $X_j$  dell'insieme assume il valore della variabile  $x_i$  osservato sul record stesso, mentre le altre sono nulle. La variabile che presenta il valore di  $x_i$  è quella che identifica il gruppo di riferimento a cui appartiene il record stesso;

- si forma un insieme di variabili  $X_j$  per ogni partizione e ogni variabile  $x_i$  in cui sono noti i totali;
- è fondamentale fare attenzione all'ordine di inserimento delle variabili al momento del lancio della procedura, in quanto il software associa la prima variabile selezionata tra i totali noti,  $TX_j$ , con la prima variabile selezionata del tipo  $X_j$ , ed analogamente accoppia la seconda, la terza (e così via) variabile  $TX_j$  selezionata con la seconda, la terza (e così via) variabile  $X_j$  selezionata<sup>9</sup>.

Una descrizione dello schema A si può ricavare dalle *tabelle 1.5 e 1.6*. In questo caso le due tabelle considerano una sola coppia partizione – variabile ausiliaria, in particolare la prima presa in considerazione, per la quale si conoscono i totali di popolazione.

Per quanto riguarda lo schema B, la sua applicazione è possibile solo quando la variabile che definisce le sottopopolazioni pianificate rientra nella definizione dei gruppi di riferimento di tutte le variabili ausiliarie.

La costruzione degli archivi di input segue, in pratica, le istruzioni illustrate nel paragrafo precedente (relative allo schema B) in cui si ha a disposizione una sola variabile ausiliaria. Le *tabelle 1.7 e 1.8* descrivono le caratteristiche dei due *data-set* di input per la prima variabile ausiliaria considerata. Per considerare anche le altre variabili ausiliarie è necessario aggiungere altrettante variabili  $TX_j$  in TOTNOTI e  $X_j$  in INP

Relativamente all'applicazione dello schema C, questa è resa possibile quando:

---

<sup>9</sup> Si ricorda che i nomi  $TX_j$  e  $X_j$  assegnati alle variabili per definire i gruppi di riferimento sono utilizzati con un puro scopo descrittivo. In realtà non esistono vincoli particolari sul tipo di nome da assegnare, e quindi, a maggior ragione, non è necessario attribuire ai nomi delle variabili Totali noti e Variabili ausiliarie un indice che le possa associare a coppie

- le modalità della variabile che definisce le sottopopolazioni pianificate sono ottenibili come combinazione di due o più variabili;
- è possibile individuare un sottoinsieme di variabili (tra quelle che definiscono le sottopopolazioni pianificate) che contribuiscono con le loro combinazioni di modalità a definire tutte le partizioni prese in considerazione dal processo di calibrazione (tale sottoinsieme di variabili definisce a sua volta delle sottopopolazioni pianificate);
- le combinazioni delle modalità delle variabili *comuni* (o un suo sottoinsieme) sono utilizzate per definire la variabile POP\_INP;
- le altre variabili che definiscono le sottopopolazioni pianificate alla base dei gruppi di riferimento, contribuiscono alla definizione delle variabili TXj in TOTINP e Xj nel *data-set* INP come avviene nello schema A e B.

La costruzione del *data-set* secondo lo schema C è illustrato sinteticamente nelle *tabelle 1.11 e 1.12*. In tali tabelle si fa riferimento ad una coppia partizione – variabile ausiliaria, in particolare la prima presa in considerazione, per la quale si conoscono i totali di popolazione.

Per ulteriori approfondimenti relativi alla costruzione dei *data-set* di input si rimanda all'*appendice A.3*. Inoltre nel *paragrafo 1.2.3* sono elencate alcune regole guida di riferimento per definire correttamente le variabili TXj e Xj.

#### **1.2.4 Scelta dello schema di costruzione del data-set TOTINP e INP**

La scelta di uno dei tre schemi per la costruzione dei *data-set* di input, deve essere operata in funzione:

- dei vincoli operativi;
- dei vantaggi e svantaggi connessi con l'efficienza computazionale del software;
- della quantità di informazioni che vengono prodotte dal software.

In particolare, per quanto riguarda il terzo punto è necessario ricordare che il software produce alcune statistiche, sotto forma di tavole e tabulati, il cui dettaglio è a livello di sottopopolazioni pianificate (cfr. *capitolo 6, Sezione I*).

Pertanto, quando si adotta lo schema A il livello di disaggregazione delle statistiche è minimo (si ha una sola sottopopolazione pianificata) mentre per gli altri due schemi le statistiche sono prodotte in una forma più articolata.

Per quanto l'efficienza computazionale, non esiste un criterio di ottimalità per indirizzare la scelta tra uno dei tre schemi o, quando si applica lo schema C, non esiste una regola per indirizzare la scelta (quando è possibile) delle due classi di variabili che definiscono le sottopopolazioni pianificate. Tuttavia, in numerose applicazioni il software si è dimostrato più efficiente quando si determina un equilibrio tra il numero delle modalità della variabile POP\_PIAN ed il numero delle variabili TXj o Xj.

Una sintesi delle linee guida per la scelta di una tra le tre alternative è fornita nella *tabella 1.13*.

**Tabella 1.13 – Vincoli, vantaggi e svantaggi dei diversi schemi di definizione del data-set di input**

Metodi di formazione del data set		
Schema A	Schema B	Schema C
Vincoli		
Non esistono vincoli sulla variabile che definisce le sottopopolazioni pianificate.	La variabile che definisce le sottopopolazioni pianificate deve essere comune a tutte le partizioni in gruppi di riferimento.	La variabile che definisce le sottopopolazioni pianificate deve essere composta da due o più variabili. Almeno una variabile o una classe di variabili che compone la variabile che definisce la sottopopolazione pianificata deve essere comune a tutte le partizioni in gruppi di riferimento.
Vantaggi		
La costruzione delle variabili di input che definiscono i gruppi di riferimento è diretta.	Nella costruzione delle variabili di input che definiscono i gruppi di riferimento si richiede solo la suddivisione tra la variabile che definisce le sottopopolazioni pianificate dalle altre variabili.  Per campioni di grandi dimensioni quando le modalità della variabile che definiscono le sottopopolazioni pianificate sono numerose può essere più efficiente dello schema A.  Massimo dettaglio delle statistiche nelle tavole e tabulati di output.	Per campioni di grandi dimensioni ed indagini multiobiettivo questa impostazione garantisce in genere una migliore efficienza computazionale quando la suddivisione delle variabili che definiscono POP_PIAN e TXj (o Xj) determina un equilibrio tra il numero delle modalità della variabile POP_PIAN ed il numero delle variabili TXj (o Xj).  Maggiore dettaglio delle statistiche nelle tavole e tabulati di output rispetto a quelle che si ottengono con lo schema A.

**Tabella 1.13 segue – Vincoli, vantaggi e svantaggi dei diversi schemi di definizione del data-set di input**

Metodi di formazione del data set		
Schema A	Schema B	Schema C
Svantaggi		
Per campioni di grandi dimensioni, ed indagini multiobiettivo si possono presentare problemi di ordine computazionale causati dal numero elevato di variabili ausiliarie TXj e Xj.  Minimo dettaglio delle statistiche nelle tavole e tabulati di output.	Per campioni di grandi dimensioni, lo schema può risultare computazionalmente meno efficiente dello schema C a causa di un eventuale numero elevato di modalità della variabile POP_PIAN.	Il metodo può richiedere alcune operazioni preventive per suddividere la variabile che definisce le sottopopolazioni pianificate in due classi di variabili.

### 1.3 Definizione delle variabili ausiliarie e della variabile “peso distanza” per calcolare i coefficienti finali di alcuni importanti stimatori campionari

Per specificare nell’ambito della teoria degli stimatori di ponderazione vincolata un particolare stimatore, si devono tenere in considerazione i seguenti aspetti:

- le variabili ausiliarie;
- il peso CK (tabella 1.2);
- la funzione utilizzata per misurare la distanza tra i coefficienti di riporto iniziali e finali.

I valori attribuiti alle variabili di input presentate nei primi due punti corrisponde, nella terminologia degli stimatori di regressione generalizzata (cfr. *appendice A.2*), alla specificazione del *tipo di modello*. L’utente in questo caso deve intervenire opportunamente sui due *data-set* TOTINP e INP.

Per quanto riguarda la funzione di distanza, questa viene selezionata attraverso le maschere introduttive al lancio della procedura (cfr. *capitolo 5*).

Di seguito si illustrano i requisiti per costruire i coefficienti finali di alcuni degli stimatori più noti in letteratura. Questi stimatori prevedono la

calibrazione sui totali calcolati a livello di unità elementare (cfr. *appendice A.2: livello del modello*).

## **Stimatore rapporto**

Di seguito sono illustrate le caratteristiche dei *data-set* di input necessarie per ottenere i coefficienti finali dei principali stimatori del rapporto.

*Stimatore Hájek* (variabile ausiliaria: numerosità di popolazione)

Nel data set TOTINP si ha:

- un solo record;
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presenta un'unica variabile TX1 con valore pari alla numerosità complessiva della popolazione di riferimento.

Relativamente a INP si ha che:

- la variabile POP\_PIAN assume un valore costante su tutto il *data-set* e uguale a quello presente in TOTINP;
- si presenta un'unica variabile ausiliaria  $X1=1$ ;
- si pone  $CK=1$ .

*Stimatore del rapporto semplice*

Nel data set TOTINP si ha:

- un solo record ;
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presenta un'unica variabile TX1 pari al totale della variabili ausiliaria su cui si basa lo stimatore.

Relativamente a INP si ha che:

- la variabile POP\_PIAN assume un valore costante su tutti i record del *data-set* e uguale a quello presente in TOTINP;
- si presenta un'unica variabile ausiliaria  $X1$  che assume i valori osservati della variabile ausiliaria su cui si basa lo stimatore (per le variabili qualitative dicotomiche il valore è pari a "1" se il record presenta

- l'attributo e "0" altrimenti);
- si pone  $CK=X_1$ ;

### *Stimatore del rapporto separato*

Questo tipo di stimatore si applica quando il disegno di campionamento è stratificato e ciascuno strato assume il ruolo di gruppo di riferimento (per approfondimento cfr. *paragrafo 1.2.1* e *appendice A.2*). Pertanto, il *data-set* di input può presentare diverse forme che variano in funzione della definizione congiunta delle variabili POP\_PIAN e dell'insieme delle variabili TXj nel *data-set* TOTINP e della variabile POP\_PIAN e delle variabili Xj nel *data-set* INP.

Bisogna inoltre distinguere il caso in cui la stratificazione è ottenuta con una variabile semplice o è il risultato di una classificazione incrociata di più variabili.

Nel primo caso il *data-set* di input può essere costruito facendo riferimento a due schemi di costruzione, denominati A e B, che possono essere utilizzati in alternativa per definire i gruppi di riferimento (per approfondimenti cfr. *paragrafo 1.2.1* e *1.2.2*). Seguendo lo schema A il data set TOTINP presenta le seguenti caratteristiche:

- esiste un solo record nel *data-set*;
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presentano tante variabili ausiliarie TXj per quanti sono gli strati del disegno. Ogni variabile TXj è associata ad uno strato. Il valore che assume ciascuna variabile TXj è pari al totale della variabile ausiliaria su cui si basa lo stimatore per il corrispondente strato identificato dalla variabile stessa (nel caso dello stimatore separato tipo Hájek la variabile TXj è pari alla numerosità complessiva della popolazione nello strato identificato dalla variabile stessa).

Il *data-set* INP è definito come segue:

- la variabile POP\_PIAN assume un valore costante su tutto il *data-set* e uguale a quello presente in TOTINP;
- si presentano tante variabili ausiliarie Xj per quanti sono gli strati del disegno. Ogni variabile ausiliaria Xj è associata ad uno strato che

- coincide con quello identificato dalla variabile  $TX_j$  in TOTINP;
- per ogni record tutte le variabili  $X_j$  sono nulle tranne quella associata allo strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (per le variabili qualitative dicotomiche il valore è pari a “1” se il record presenta l’attributo e “0” altrimenti; nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone  $CK$  pari al valore della variabile ausiliaria su cui si basa lo stimatore e che si osserva sul record.

Per quanto riguarda l’impostazione dello schema B, il *data set* TOTINP deve essere costruito come segue:

- esistono tanti record per quanti sono gli strati del disegno. Ciascun record identifica uno strato;
- la variabile POP\_PIAN presenta tante modalità per quanti sono gli strati del disegno. Ciascun record presenta una modalità diversa della variabile;
- esiste una sola variabile  $TX_1$  che assume un valore pari al totale della variabile ausiliaria su cui si basa lo stimatore per il corrispondente strato identificato dal record (nel caso dello stimatore separato tipo Hájek la variabile  $TX_j$  è pari alla numerosità complessiva della popolazione nello strato identificato dal record).

Se TOTINP viene costruito con lo schema B, allora INP deve essere definito come segue:

- ciascun record presenta per la variabile POP\_PIAN la stessa modalità presente nel *data-set* TOTINP che identifica lo strato a cui appartiene il record stesso;
- esiste una sola variabile  $X_1$  che assume il valore della variabile ausiliaria su cui si basa lo stimatore (per le variabili qualitative dicotomiche il valore è pari a “1” se il record presenta l’attributo e “0” altrimenti; nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone  $CK=X_1$ .

Quando la variabile di stratificazione è ottenuta dalla combinazione delle modalità di più variabili (variabile combinata) è possibile anche definire i due *data-set* secondo una terza alternativa denominata schema C (per



approfondimenti cfr. *paragrafo 1.2.1 e 1.2.2*). In questo caso, le variabili originali di stratificazione sono divise in due classi complementari. La prima classe definisce il numero di valori che assume la variabile POP\_PIAN, attraverso la combinazione delle modalità delle variabili ad esso appartenenti; mentre la seconda classe determina, mediante la classificazione incrociata delle modalità delle variabili in esso contenute, il numero di variabili ausiliarie TXj e Xj. Si ha quindi che in TOTINP:

- esistono tanti record per quante sono le combinazioni delle modalità della prima classe di variabili. Ciascun record identifica una combinazione;
- la variabile POP\_PIAN presenta tante modalità per quante sono le combinazioni delle modalità della prima classe di variabili. Ciascun record presenta una modalità diversa della variabile;
- si presentano tante variabili ausiliarie TXj per quante sono le combinazioni delle modalità della seconda classe di variabili. Ogni variabile TXj è associata ad una combinazione. Il valore che assume ciascuna variabile TXj per un dato record è pari al totale della variabile ausiliaria su cui si basa lo stimatore per il corrispondente strato identificato dalla combinazione della modalità della variabile POP\_PIAN e della modalità che identifica la stessa TXj (nel caso dello stimatore separato tipo Hájek la variabile TXj è pari alla numerosità complessiva della popolazione nello strato).

Secondo lo schema C il data set INP viene definito come segue:

- ciascun record presenta per la variabile POP\_PIAN la stessa modalità presente nel data set TOTINP che identifica la combinazione delle modalità della prima classe di variabili che presenta il record stesso;
- si presentano tante variabili ausiliarie Xj pari al numero delle combinazioni delle modalità della seconda classe di variabili;
- per ogni record tutte le variabili Xj sono nulle tranne quella associata alla combinazione di modalità della seconda classe di variabili che presenta il record stesso. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (per le variabili qualitative dicotomiche il valore è pari a “1” se il record presenta l’attributo e “0” altrimenti; nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);

- si pone  $CK$  pari al valore della variabile ausiliaria su cui si basa lo stimatore e che si osserva sul record.

È importante ricordare che in tutti e tre gli schemi per ogni coppia  $TX_j$  e  $X_j$  le variabili devono identificare lo stesso strato o combinazione di modalità.

Infine si rimanda al *paragrafo 1.2.4.* per avere alcuni suggerimenti sulla scelta di uno dei tre schemi, ricordando che la scelta tra una delle tre alternative dipende anche da come sono organizzati i dati di input prima di iniziare le operazioni di costruzione dei *data-set* TOTINP e INP.

### *Stimatore del rapporto combinato*

Per questo tipo di stimatore, il *data-set* deve presentare i seguenti requisiti nel *data-set* TOTINP:

- esiste un solo record nel *data-set*;
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presenta un'unica variabile  $TX1$  pari al totale della variabile ausiliaria su cui si basa lo stimatore;

Relativamente a INP si ha che:

- la variabile POP\_PIAN assume un valore costante su tutti i record e uguale a quello presente in TOTINP;
- si presenta un'unica variabile ausiliaria  $X1$  che assume i valori osservati della variabile ausiliaria su cui si basa lo stimatore;
- si pone  $CK=X1$ .

### *Stimatore rapporto post-stratificato*

Per quanto riguarda il *data-set* TOTINP si ha:

- un solo record nel *data-set*
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presentano tante variabili ausiliarie  $TX_j$  per quanti sono i post-strati. Ogni variabile  $TX_j$  è associata ad un post-strato. Il valore che assume ciascuna variabile  $TX_j$  è pari al totale della variabili ausilia-

ria su cui si basa lo stimatore per il corrispondente post-strato identificato dalla variabile stessa.

Il *data-set* INP è definito come segue:

- la variabile POP\_PIAN assume un valore costante su tutto il *data-set* e uguale a quello presente in TOTINP;
- si presentano tante variabili ausiliarie  $X_j$  per quanti sono i post-strati del disegno. Ogni variabile ausiliaria  $X_j$  è associata ad un post-strato che coincide con quello identificato dalla variabile  $TX_j$  in TOTINP;
- per ogni record tutte le variabili  $X_j$  sono nulle tranne quella associata al post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore.
- si pone CK pari al valore della variabile ausiliaria su cui si basa lo stimatore e che si osserva sul record.

#### *Stimatore rapporto post-stratificato separato*

Lo stimatore del rapporto post-stratificato separato richiede la conoscenza dei totali della variabile ausiliaria su particolari sottopopolazioni, denominate gruppi di riferimento, definite all'interno di ciascuno strato (cfr. *appendice A.2*). Ciascun gruppo di riferimento è identificato da una modalità  $d$  ( $d=1, \dots, D$ ) di una variabile qualitativa. In particolare per lo stimatore in questione, tale variabile è definita attraverso la combinazione di due variabili: la prima variabile identifica gli strati del disegno; la seconda variabile identifica i post-strati all'interno di ciascuno strato. Per indicare al software quali sono i gruppi di riferimento da considerare l'utente può seguire tre diverse alternative. In questo paragrafo sono descritte due alternative denominate schema A, schema B. Per quanto riguarda la terza alternativa, denominata schema C, si rimanda al *paragrafo 1.2.2*.

Considerando lo schema A il *data-set* TOTINP si deve avere:

- un solo record nel *data-set*;
- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si presentano tante variabili ausiliarie  $TX_j$  per quante sono le combinazioni strati per post-strati. Ogni variabile  $TX_j$  è associata ad una di queste combinazioni. Il valore che assume ciascuna variabile  $TX_j$

è pari al totale della variabile ausiliaria su cui si basa lo stimatore per la corrispondente combinazione strato e post-strato identificato dalla variabile stessa.

Il *data-set* INP è definito come segue:

- la variabile POP\_PIAN assume un valore costante su tutto il *data-set* e uguale a quello presente in TOTINP;
- si presentano tante variabili ausiliarie  $X_j$  per quante sono le combinazioni strati per post-strati. Ogni variabile ausiliaria  $X_j$  è associata ad una combinazione strato per post-strato e coincide con quella identificata dalla variabile  $TX_j$  in TOTINP;
- per ogni record tutte le variabili  $X_j$  sono nulle tranne quella associata alla combinazione strato per post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (per le variabili qualitative dicotomiche il valore è pari a “1” se il record presenta l’attributo e “0” altrimenti);
- si pone CK pari al valore della variabile ausiliaria su cui si basa lo stimatore e che si osserva sul record.

Una seconda impostazione dei due *data-set* è data dallo schema B. In questo caso il *data-set* TOTINP deve essere costruito come segue:

- esistono tanti record per quanti sono gli strati del disegno. Ciascun record identifica uno strato;
- la variabile POP\_PIAN presenta tante modalità per quanti sono gli strati del disegno. Ciascun record presenta una modalità diversa della variabile;
- esistono tante variabili  $TX_j$  per quanti sono i post-strati all’interno di uno strato. Ciascuna  $TX_j$  assume un valore pari al totale della variabile ausiliaria su cui si basa lo stimatore per il corrispondente incrocio tra la modalità della variabile POP\_PIAN e il post-strato identificato dalla stessa  $TX_j$ .

Se TOTINP viene costruito con lo schema B, allora INP deve essere definito come segue:

- ciascun record presenta per la variabile POP\_PIAN la stessa modalità presente nel *data-set* TOTINP che identifica lo strato a cui appartiene il record stesso;
- esistono tante variabili  $X_j$  per quanti sono i post-strati all’interno di

uno strato. Ciascuna variabile ausiliaria assume valore nullo tranne quella  $X_j$  che, combinata con la modalità della variabile POP\_PIAN che presenta il record, definisce il post-strato a cui appartiene il record stesso. In questo caso  $X_j$  ha come valore quello della variabile ausiliaria su cui si basa lo stimatore e che è stato osservato sullo stesso record;

- si pone CK pari al valore della variabile ausiliaria su cui si basa lo stimatore e che si osserva sul record.

E' importante ricordare nei due schemi descritti per ogni coppia  $TX_j$  e  $X_j$  le variabili devono identificare lo stesso strato o combinazione di modalità.

Infine si rimanda al *paragrafo 1.2.4.* per avere alcuni suggerimenti sulla scelta di uno fra gli schemi A, B e lo schema C (presentato nel *paragrafo 1.2.2.*).

#### *Stimatore rapporto post-stratificato combinato*

I coefficienti finali dello stimatore del rapporto post-stratificato si ottengono seguendo le stesse istruzioni suggerite per lo stimatore del rapporto separato secondo lo schema A. In questo caso, i post-strati assumono il ruolo che hanno gli strati per il precedente stimatore.

### **Stimatore raking**

Gli stimatori raking utilizzano i totali di popolazione di una sola variabile ausiliaria per sottopopolazioni, denominate gruppi di riferimento, appartenenti a diverse partizioni distinte della popolazione obiettivo (cfr. *appendice A.2*). In particolare lo stimatore ratio raking considera due partizioni in gruppi di riferimento, mentre lo stimatore raking generalizzato estende la calibrazione a totali per più di due partizioni in gruppi di riferimento della popolazione obiettivo.

#### *Stimatore ratio raking*

Siano  $Q_1$  e  $Q_2$  il numero di gruppi di riferimento di una popolazione definiti, rispettivamente, sulla base delle modalità assunte dalle variabili ausiliarie  $v_1$  e  $v_2$ . Il *data-set* TOTINP presenta le seguenti caratteristiche:

- esiste un solo record;

- la variabile POP\_PIAN assume un valore scelto arbitrariamente dall'utente;
- si assegna un numero progressivo ai gruppi di riferimento relativi alle due partizioni; i gruppi di riferimento della prima partizione hanno un numero progressivo da 1 a  $Q_1$ , mentre quelli della seconda partizione hanno numeri progressivi che vanno da  $Q_1+1$  a  $Q_1+Q_2$ ;
- a ciascuno dei  $Q_1+Q_2$  gruppi di riferimento si associa una variabile  $TX_j$  con  $j=1, \dots, Q_1+Q_2$ ;
- il valore assunto da ciascuna  $TX_j$  è pari al totale osservato della variabile ausiliaria sul corrispondente gruppo di riferimento.

Per definire il *data-set* INP si deve avere che:

- la variabile POP\_PIAN assume un valore costante su tutto il *data-set* e uguale a quello presente in TOTINP;
- a ciascuno dei  $Q_1+Q_2$  gruppi di riferimento si associa una variabile  $X_j$  con  $j=1, \dots, Q_1+Q_2$ . Ad ogni coppia  $TX_j, X_j$  corrisponde lo stesso gruppo di riferimento;
- ciascuna variabile  $X_j$  assume valore nullo tranne quella che corrisponde al gruppo di riferimento a cui appartiene il record. In questo caso  $X_j$  è posta pari a "1" (per ogni record sono presenti due variabili  $X_j$  pari a "1");
- la variabile  $CK=1$  per tutti i record del *data-set*.

### *Stimatore raking generalizzato*

L'impostazione dei *data-set* di input per ottenere i coefficienti di riporto finali dello stimatore è facilmente ricavabile da quanto illustrato per lo stimatore ratio raking. Questo stimatore generalizza il precedente considerando un insieme di variabili qualitative ( $V_1, \dots, V_G$ ) che definiscono  $G(>2)$  partizioni in gruppi di riferimento.

Riprendendo la simbologia e le regole descritte per lo stimatore ratio raking è, pertanto, necessario definire  $Q_1 + \dots + Q_G$  variabili  $TX_j$  e  $X_j$ .

La costruzione dei *data-set* avviene in analogia a quanto descritto per lo stimatore ratio raking.

## Stimatori di regressione generalizzata e di ponderazione vincolata

Gli stimatori di regressione generalizzata o la più ampia classe degli stimatori di ponderazione vincolata utilizzano i totali noti per sottopopolazioni, denominate gruppi di riferimento, appartenenti a partizioni distinte della popolazione obiettivo (cfr. *appendice A.2*). Per tale struttura dei totali noti la costruzione dei *data-set* di input può seguire diverse alternative. Queste sono descritte nei *paragrafi 1.2.2 e 1.2.3*.

Un esempio dettagliato che descrive la costruzione dei due *data-set* di input è illustrato nell'*appendice A.3*.

### 1.4. Scelta della funzione di distanza per definire i coefficienti finali di alcuni stimatori campionari

Per indicare al software lo stimatore che si intende utilizzare, oltre a definire i due *data-set* di input TOTINP e INP (cfr. *paragrafo 1.2*), è necessario selezionare dalle maschere di avvio una funzione di distanza. Si ricorda infatti, che il software definisce gli stimatori nell'ambito della classe degli stimatori di ponderazione vincolata i quali, a loro volta, si basano sulla scelta di una funzione di distanza (per approfondimenti cfr. *appendice A.1*).

Il software implementa sette funzioni di distanza come indica la *tabella 1.14*.

---

**Tabella 1.14 – Funzioni di distanza implementate dal software**

Funzione di distanza
1 – Logaritmica troncata
2 – Logaritmica
3 – Euclidea (o Lineare)
4 – Euclidea troncata
5 – Hellinger
6 – Minima Entropia
7 – Chi quadrato

---

Nella tabella seguente sono presentate le distanze da utilizzare per ottenere i coefficienti finali di alcuni tra i principali stimatori noti in letteratura.

**Tabella 1.15 – Funzioni di distanza che definiscono alcuni stimatori**

<i>Stimatore</i>	<i>Funzione di distanza</i>
Hàjek	Euclidea
Rapporto semplice/separato/combinato	Euclidea
Rapporto post-stratificato/separato/combinato	Euclidea
Ratio raking/raking generalizzato	Logaritmica
Regressione generalizzata	Euclidea

Per quanto riguarda gli stimatori ratio raking / raking generalizzato e di regressione generalizzata, l'uso delle funzioni di distanza utilizzate può comportare alcuni problemi relativi ai valori assunti dai coefficienti finali di output, i quali potrebbero essere molto elevati o negativi. Per evitare questi problemi o incongruenze si può ricorrere alla funzione di distanza logaritmica troncata per il ratio raking / raking generalizzato e alla funzione euclidea troncata per lo stimatore di regressione generalizzata. Tali distanze limitano infatti i coefficienti finali all'interno di un *range* prefissato dall'utente. Il ricorso ad una funzione di distanza troncata può essere necessario anche per stimatori di ponderazione vincolata più complessi che non rientrano tra quelli specificati nella tabella. E', tuttavia, opportuno ricordare che l'uso di funzioni di distanza troncate può determinare problemi di convergenza dell'algoritmo iterativo per il calcolo dei coefficienti finali qualora il *range* imposto sui coefficienti fosse troppo ristretto.

Per la descrizione dell'uso delle maschere di lancio del software per selezionare la funzione di distanza si veda il *capitolo 5* della *Sezione I*.

### **1.5 Utilizzo del software per il trattamento delle mancate risposte totali (unità non rispondenti)**

Il software assume che l'insieme dei record del *data-set* INP rappresenti il campione completo definito in fase di pianificazione dell'indagine. Tale



ipotesi, tuttavia, soprattutto per campioni di grandi dimensioni non è in pratica realistica, in quanto spesso alcune unità selezionate nel campione non partecipano all'indagine (unità non rispondenti). La presenza di unità non rispondenti, dette anche mancate risposte totali (che non sono comprese nel *data-set* INP), se non è tenuta in giusta considerazione nel processo di stima, può produrre seri effetti distorsivi sulle stime finali.

Generalmente per eliminare o ridurre la distorsione degli stimatori, dovuta al problema delle mancate risposte totali, si applicano procedure di aggiustamento dei coefficienti di riporto iniziali delle unità rispondenti per compensare la presenza di unità non rispondenti. In estrema sintesi, la procedura corregge i coefficienti iniziali in modo tale che con essi si possano riprodurre, con il solo campione dei rispondenti, i totali (noti da fonti esterne) per diverse sottopopolazioni di alcune variabili ausiliarie correlate con il fenomeno della mancata risposta totale. Un tale approccio per la correzione dei coefficienti di riporto prevede la definizione di un modello di dipendenza della probabilità di mancata risposta con le variabili utilizzate per correggere la mancata risposta totale. Se questo modello è valido si ha una riduzione della distorsione delle stime dei parametri di interesse causata dall'osservazione del solo campione di rispondenti. Per la scelta delle variabili e delle sottopopolazioni definite per il trattamento delle mancate risposte totali si rimanda alla letteratura (si veda ad esempio Little, 1986).

Il processo di correzione per mancata risposta totale si prefigura come un processo di calibrazione in cui le sottopopolazioni sulle quali sono noti i totali, sono denominate celle (o classi) di aggiustamento (o di omogeneità). Tali celle di aggiustamento assumono, quindi, il ruolo dei gruppi di riferimento degli stimatori di ponderazione vincolata. Di seguito sono elencate alcune procedure di aggiustamento dei coefficienti iniziali applicabili con il software.

**Tabella 1.16 – Alcune procedure di correzione dei coefficienti iniziali di riporto per mancata risposta totale e relative impostazioni dei data-set TOTINP e INP**

<i>Caratteristiche dei Coefficienti Corretti (CC)</i>	<i>Impostazione dei data-set TOTINP e INP come:</i>
I CC riproducono la numerosità della popolazione.	Stimatore di Hájek
I CC riproducono la frequenza totale per ciascuna classe di aggiustamento.	Stimatore del rapporto post-stratificato tipo Hájek
I CC riproducono la frequenza totale per ciascuna classe di aggiustamento di due distribuzioni distinte.	Stimatore ratio raking
I CC riproducono la frequenza totale per ciascuna classe di aggiustamento di tre o più distribuzioni distinte.	Stimatore raking generalizzato
I CC riproducono il totale di popolazione della variabile ausiliaria usata per correggere.	Stimatore del rapporto
I CC riproducono il totale di popolazione della variabile ausiliaria usata per correggere per ciascuna classe di aggiustamento.	Stimatore del rapporto post-stratificato
I CC riproducono il totale di popolazione per una o più variabili ausiliarie usate per correggere, per ciascuna classe di aggiustamento di due o più distribuzioni distinte.	Stimatore di regressione generalizzata

Solo dopo aver compiuto questa operazione, si procede alla fase di calibrazione considerando i coefficienti corretti per mancata risposta totale. Nel caso in cui si usano le stesse variabili ausiliarie nella fase di correzione e in quella di calibrazione e quando per ciascuna variabile le celle di aggiustamento coincidono con i gruppi di riferimento, lo stimatore di ponderazione vincolata corregge anche la mancata risposta totale.

Per ripercorrere con il software il caso più generale del processo di stima in due passi, è necessario eseguire il lancio della procedura informatica due volte.

## **Passo 1 – correzione per mancata risposta**

Sul *data-set* TOTINP e INP si definiscono le celle di aggiustamento per mancata risposta seguendo le stesse istruzioni necessarie per definire i gruppi di riferimento (*cfr. paragrafo 1.2*). Le variabili interessate sono POP\_PIAN e le variabili di tipo TXj e Xj. Inoltre nel *data-set* INP sono presenti anche le variabili Xj che devono essere utilizzate per la fase di calibrazione (passo 2). Di seguito sono fornite le principali istruzioni per costruire i due *data-set* di input per realizzare questo passo:

- La variabile POP\_PIAN e le variabili TXj in TOTINP e Xj in INP siano definite per individuare le celle di aggiustamento (si veda per approfondimenti il *paragrafo 1.2* considerando che in questo caso le celle di aggiustamento rappresentano i gruppi di riferimento);
- è presente un secondo insieme di variabili Xj in INP che nel passo 2 (passo di calibrazione) saranno utilizzate per calibrare i coefficienti corretti per mancata risposta totale. La definizione di tali variabili dipende dalle variabili di calibrazione e dai gruppi di riferimento che saranno usati nel passo 2 (*cfr. paragrafo 1.2*);
- l'impostazione dei due *data-set* per un determinato tipo di correzione si completa considerando un particolare processo di calibrazione sulle variabili ausiliarie considerate per correggere la mancata risposta. Alcune correzioni sono indicate nella *tabella 1.16* (si veda anche il *paragrafo 1.2 e 1.3*);
- la variabile COEF indica il coefficiente diretto

## **Passo 2 – calibrazione dei coefficienti.**

Eseguito il primo lancio della procedura si deve effettuare la calibrazione dei coefficienti corretti per mancata risposta totale.

A tal fine bisogna eseguire i punti seguenti:

- si considera il *data-set* INP ottenuto come output dell'applicazione del software al passo 1. In tale *data-set* compare la variabile COEF-FIN che è il coefficiente iniziale corretto per mancata risposta;
- la variabile COEFFIN diventa il coefficiente di riporto iniziale per il secondo lancio della procedura. A scopo illustrativo si può rinominare la variabile affinché non venga cancellata alla fine del secondo lancio della procedura (che produce una nuova variabile COEF-

FIN) chiamando COEFFIN con il nome COEF (il questo caso si elimina il coefficiente iniziale) proprio per indicare che tale variabile assume il ruolo di coefficiente di riporto iniziale per il secondo lancio della procedura;

- si costruisce un nuovo *data-set* TOTINP, in cui la variabile POP\_PIAN deve tenere conto dei gruppi di riferimento dello stimatore adottato (cfr. per approfondimenti il *paragrafo 1.2*). In particolare, deve esistere una corrispondenza tra la generica variabile TX<sub>j</sub> e una variabile X<sub>j</sub> del secondo gruppo di variabili X<sub>j</sub> definite nel *data-set* INP per la calibrazione;
- la variabile che al primo passo è stata chiamata POP\_PIAN nel *data-set* INP deve essere sostituita. A scopo illustrativo si elimini tale variabile e se ne consideri una nuova con lo stesso nome usato in TOTINP (nell'esempio POP\_PIAN) tale che definisca con le variabili X<sub>j</sub> i gruppi di riferimento del modello (nella pratica la variabile che è stata chiamata POP\_PIAN può assumere un qualsiasi nome. E' importante che tale nome sia lo stesso nei due *data-set* TOTINP e INP e sia costruita in modo tale da definire, in questo secondo passo, i gruppi di riferimento);

Alla fine del secondo lancio della procedura la variabile di output COEFFIN rappresenta i coefficienti finali di riporto corretti per mancata risposta totale.



## 2. L'output del software

**Sintesi:** Il software produce alcuni data-set di output, file ascii e file excel, scritti sulla cartella di output scelta dall'utente.

I data-set di output sono i seguenti<sup>10</sup>:

*Data-set creato per memorizzare parametri di input (paragrafo 2.1)*

- *SAVESTIME*

*Data-set creati per memorizzare gli errori rilevati sull'input (paragrafo 2.2)*

- *NOTI\_MISS, CODICI\_DOPPI, CSENZAT, MISSING, VUOTI,*

*Data-set contenenti i pesi finali (paragrafo 2.3)*

- *PESIFIN, STAT, STIMEDIR. STIMEFIN,*

### 2.1 Il data-set dei parametri di input

Nel *paragrafo 5.2.3 della Sezione I* viene descritto come selezionare le variabili di input tramite i parametri della procedura. Ciò è possibile in quanto il software, per ciascuna elaborazione, scrive nella cartella di output il *data-set* SAVESTIME, indispensabile per attivare la funzione “Parametri attivi”.

In *figura 5.1* è possibile vedere un esempio di *data-set* SAVESTIME: il *data-set* è caratterizzato da due soli campi “*descr*” e “*parametro*” e il software

---

<sup>10</sup> La cartella di output scelta dall'utente corrisponde alla libreria “outtime”. Se, ad esempio, l'utente sceglie la cartella c:\utente - prendendo in considerazione il data-set di output STAT - la procedura crea il data-set sas di output “outtime.stat” che corrisponde al file c:\utente\STAT.sas7bdat (data-set sas v.8) registrato nella cartella c:\utente. Per semplificare l'esposizione successiva si farà riferimento ai data-set solo con il nome, senza l'estensione del file o la libreria di riferimento

scrive in automatico le 13 righe del *data-set*, memorizzando le scelte precedentemente fatte dall'utente.

**Figura 2.1: Il data-set SAVESTIME**

	descr	parametro
1	INPUT1 - dsn dati campionari	G:\esempi\Inpstime.Input
2	INPUT2 - dsn totali	G:\esempi\Inpstime.Noti
3	Peso distanza	CK
4	Variabili ausiliarie	X1 X2
5	Peso diretto	COEF
6	Pop. pianif. per stimatore INPUT1	POPOL
7	Codice identif. unità campionaria	CODICE
8	Cod. pop. pianif. per stimatore	.
9	Pop. pianif. per stimatore INPUT2	POPOL
10	Totali variabili ausiliarie	TX1 TX2
11	Funzione di distanza	1
12	Coeff.molt.val.minimo di L	0.5
13	Coeff.molt.val.massimo di U	1.5

## 2.2 Gli errori rilevati sul data-set di input

Il software Genesees è predisposto al controllo automatico di alcuni errori rilevati sui *data-set* di input. Generalmente il software ferma l'elaborazione e scrive i *data-set* contenenti gli errori riscontrati. In un solo caso il software scrive un *data-set* ma non blocca l'elaborazione: infatti, come si vedrà più avanti, la scrittura del *data-set* VUOTI non è associata al blocco dell'elaborazione, in quanto, più che un errore dell'utente, il software potrebbe aver riscontrato un caso di mancata risposta, relativo ad una specifica partizione. Per approfondimenti, cfr. *paragrafo 5.1.1 Sezione I*.

### 2.2.1 I controlli sul data-set dei totali noti e la scrittura del data-set NOTI\_MISS

Nel caso di valori mancanti assunti dalla variabile "Popolazioni pianificate utilizzate per lo stimatore" nel *data-set* dei totali noti, il software si blocca e manda una segnalazione di errore.

La variabile "Popolazioni pianificate utilizzate per lo stimatore" contenente i valori mancanti è scritta nel *data-set* NOTI\_MISS.

**Tabella 2.2 – Data-set NOTI\_MISS**

Nome variabile	Significato variabile nel <i>data-set</i>
DOMINIO	Codice della popolazione pianificata utilizzata per lo stimatore
TX1	Totale noto variabile ausiliaria X1
.....	.....
TXn	Totale noto variabile ausiliaria Xn

### **2.2.2 I controlli sul data-set dei dati campionari e la scrittura del data-set MISSING e del data-set CODICI DOPPI**

Il software si blocca e manda una segnalazione di errore se ci sono valori mancanti assunti dalla variabile “Codice”. La variabile “Codice” contenente i valori mancanti è scritta nel *data-set* MISSING.

Il software controlla che nel *data-set* dei dati campionari ciascuna unità campionaria sia identificata da un valore diverso della variabile “Codice”. Non possono perciò esistere record identificati da uno stesso valore della variabile “Codice”, che rappresenta una chiave univoca. In questo caso il software crea il *data-set* CODICI\_DOPPI.

**Tabella 2.3 Data-set CODICI-DOPPI**

Nome variabile	Significato variabile nel <i>data-set</i>
Codice	Valore della variabile “Codice” che risulta duplicata

### **2.2.3 Il controllo della corrispondenza tra i valori dei due data-set di input e la scrittura dei data-set CSENZAT e VUOTI**

Come descritto nel *paragrafo 5.1, Sezione I*, in entrambi i *data-set* deve essere definita la variabile “Popolazioni pianificate utilizzate per lo stimatore”.

Può accadere che per una certa popolazione pianificata di cui è noto il totale (e che dunque esiste nel *data-set* dei totali noti) non vi sia alcuna unità corrispondente nel *data-set* dei dati campionari. **In questo caso il software non viene bloccato, ma procede con l’elaborazione.** E’ infatti probabile che non si tratti di un errore, ma che nessuna delle unità selezionate nel campione appartenga a tale popolazione (caso di mancata risposta). Per segnalare tali casi, viene creato il *data-set* VUOTI .



**Tabella 2.4 – Data-set VUOTI**

Nome variabile	Significato variabile nel <i>data-set</i>
DOMINIO	Codice della popolazione pianificata utilizzata per lo stimatore che appare nel <i>data-set</i> dei totali noti ma non presenta valori nel <i>data-set</i> dei dati campionari
TX1	Totale noto variabile ausiliaria X1
.....	.....
TXn	Totale noto variabile ausiliaria Xn

Quando il software verifica che nel *data-set* dei dati campionari, una unità campionaria appartiene ad una popolazione pianificata che non esiste nel *data-set* dei totali noti, l'elaborazione viene bloccata. Per segnalare tali casi, viene creato il *data-set* CSENZAT. E' possibile vedere un esempio di tale *data-set* in figura 2.2.

**Figura 2.2 - Data-set CSENZAT**

Columns		
Column Name	Type	Length
123. 67.. ck	Number	8
123. 67.. coef	Number	8
Aa dominio	Text	15
123. 67.. codice	Number	8
123. 67.. x1	Number	8
123. 67.. x2	Number	8

E' da osservare che tale *data-set* viene scritto anche nel caso in cui una popolazione assume un valore mancante, in quanto il software non riconosce il corrispondente valore nel *data-set* dei totali noti.

## 2.3 I Data-set contenenti i pesi finali

I *data-set* sas creati dalla procedura sono diversi, tuttavia si descrivono di seguito solamente quelli che possono essere utili all'utente per applicazioni successive, tali *data-set* sono:

- PESIFIN;

- STAT ;
- STIMEDIR;
- STIMEFIN;

È da rilevare che - dopo l'elaborazione - il *data-set* di input contenente i dati campionari include in aggiunta una variabile, corrispondente ai pesi finali calcolati.

I *data-set* di cui sopra contengono tutte variabili numeriche e sono formati nel modo di seguito illustrato.

PESIFIN è il *data-set* più importante in quanto contiene il peso finale  $w_k$ , il peso diretto  $d_k$  ed il coefficiente di correzione  $\gamma_k$  per ciascuna delle  $n$  unità rispondenti all'indagine. Esso è composto di  $n$  osservazioni e 6 variabili indicate nell'ordine con i nomi CODICE, CONTA, DOMINIO, D, F e W che rappresentano rispettivamente, per ciascuna delle  $n$  unità campionarie: il codice,  $k$  (identificativo dell'unità), il numero progressivo associato a ciascun dominio; il codice di dominio di studio cui l'unità appartiene, il peso diretto  $d_k$ , il coefficiente di correzione  $\gamma_k$  ed il peso finale  $w_k$ . L'utente, al fine di produrre le tabelle di pubblicazione, deve accoppiare il *data-set* PESIFIN al file contenente gli  $n$  record con le variabili rilevate nell'indagine campionaria. L'accoppiamento deve avvenire sulla base del codice identificativo dell'unità. Le stime vengono, infine ottenute sulla base della formula (3) dell'app. A.1. Il *data-set* ha la seguente struttura:

---

**Tabella 2.5 - Struttura del data-set PESIFIN**

<i>Nome Variabile</i>	<b>Contenuto del campo</b>
Dominio	Codice di dominio oggetto di studio
Codice	Codice identificativo dell' unità
W	Peso finale ( $w_k$ )
D	Peso diretto ( $d_k$ )
F	Correttore del Peso diretto ( $\gamma_k$ )

---

STAT è il *data-set* contenente i valori delle statistiche sui pesi diretti, sui coefficienti di correzione dei pesi diretti e sui pesi finali per ciascuno degli A domini di studio. Esso, inoltre, contiene per ogni dominio alcune informazioni sull'andamento della procedura iterativa. Esso è formato da A osservazioni e 25 variabili, di cui le più importanti indicano:

- CONTA, che numera progressivamente i domini di studio;
- L, U e MAXITER , che denotano rispettivamente i valori assegnati dall'utente ai parametri L, U e il valore di default assegnato dalla procedura al numero massimo di iterazioni;
- MAXI, MAXD, MAXF MAXW,
- MIND, MINF MINW,
- SUMD, SUMW, SUMF,
- VARD, VARW, VARF,
- MEAND, MEANW, MEANF,
- CVD, CVW, CVF;

che indicano rispettivamente:

- il valore massimo;
- il valore minimo;
- la somma;
- la varianza;
- la media;
- il coefficiente di variazione;  
dei pesi diretti (ultima lettera D), dei pesi finali (ultima lettera W) e dei coefficienti di correzione (ultima lettera F);
- ITER, che è il numero di iterazioni realizzato;
- R2 che indica il numero di unità campionarie nel dominio;
- C2 che è il numero J dei vincoli definiti dall'utente.

I *data-set* STIMEDIR, STIMEFIN hanno una struttura simile e sono composti ciascuno da (**A x J**) osservazioni.

STIMEDIR, in particolare, contiene i valori dei totali noti e delle corrispondenti stime campionarie dirette. Ciascuna osservazione è relativa alla specifica concatenazione di un dominio di studio con la generica variabile  $x_j$ . Le variabili del *data-set* sono indicate con i nomi: CONTA, TOTALE, TX, TXD, DIFFD, G, che denotano, rispettivamente: il numero progressivo associato a ciascun dominio di studio (CONTA), il codice  $j$  relativo alla variabile  $x_j$  (TOTALE), il totale noto (TX), la stima diretta (TXD), la differenza assoluta tra totale noto e corrispondente stima diretta (DIFFD), il valore del moltiplicatore di Lagrange (G) relativo alla funzione di distanza utilizzata.

STIMEFIN contiene, invece, i valori dei totali noti e delle corrispondenti stime campionarie finali. Ciascuna osservazione è relativa alla specifica concatenazione di un dominio di studio con la generica variabile  $x_j$ . Le variabili del *data-set* sono indicate con i nomi: CONTA, TOTALE, TX, TXW, DIFFW, G, che denotano, rispettivamente: il numero progressivo associato a ciascun dominio di studio (CONTA), il codice  $j$  relativo alla variabile  $x_j$  (TOTALE), il totale noto (TX), la stima diretta (TXD), la differenza assoluta tra totale noto e corrispondente stima finale (DIFFD), il valore del moltiplicatore di Lagrange (G) relativo alla funzione di distanza utilizzata.

# APPENDICI

## A.1. Stimatori di ponderazione vincolata

### A.1.1. Premessa

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

- (i) le *incognite* da determinare sono i pesi finali;
- (ii) il *sistema di vincoli* assicura il rispetto della condizione di uguaglianza tra i totali noti (delle variabili ausiliarie) e le corrispondenti stime campionarie, calcolate sulla base dei pesi finali;
- (ii) la *funzione obiettivo* è una funzione di distanza tra i pesi finali incogniti e i *pesi diretti*, ottenuti come reciproco delle probabilità d'inclusione delle unità campionarie.

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

Tutti gli stimatori adottati nella pratica delle indagini campionarie su larga scala sono casi particolari dello stimatore di ponderazione vincolata, ottenuti definendo in modo opportuno la funzione di distanza impiegata. Nel

caso in cui si utilizza la funzione di distanza euclidea, lo stimatore di ponderazione vincolata che si ottiene è uguale a quello di regressione generalizzata. Per quanto concerne gli altri stimatori di ponderazione vincolata, usualmente utilizzati nella pratica delle indagini campionarie (derivanti da funzioni di distanza per le quali sono verificate condizioni di regolarità piuttosto generali, Deville e Särndal, 1992 §), è possibile dimostrare la loro tendenza asintotica allo stimatore di regressione generalizzata. Quest'ultima caratteristica riveste una notevole importanza sia dal punto di vista pratico che dal punto di vista teorico in quanto, nelle indagini su larga scala basate su campioni di notevole ampiezza, tutti gli stimatori di ponderazione vincolata assumono le medesime proprietà dello stimatore di regressione generalizzata, ossia:

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

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

far riferimento al problema di minimo vincolato sottostante alla costruzione degli stimatori di ponderazione vincolata; in tale ottica, sembra più logico parlare di: (i) *gruppo di riferimento dello stimatore*, (ii) *livello dello stimatore*, e (iii) *tipo di stimatore*; questi ultimi concetti saranno brevemente illustrati nel seguito.

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

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

I *gruppi* rappresentano una partizione completa della popolazione.

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

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

### **A.1.2 Caratteristiche generali degli stimatori di ponderazione vincolata**

Sia  $U$  una popolazione finita di  $N$  elementi, che indichiamo come  $U = \{1, \dots, k, \dots, N\}$ , e sia  $s$  un campione casuale di  $n$  elementi estratto da  $U$ , che indichiamo come  $s = \{1, \dots, k, \dots, n\}$ , mediante un disegno di campionamento che genera l'*universo dei campioni*  $S$  (ossia l'insieme di tutti i possibili campioni estraibili mediante il disegno in parola) ed assegna al generico campione  $s$  la probabilità  $p(s)$  di essere estratto, dove  $\sum_{s \in S} p(s) = 1$ .



Con riferimento alla generica unità  $k \in U$ , indichiamo quindi con:  
 $\pi_k = \sum_{s \in S(k)} p(s)$ , la probabilità di inclusione nel campione dell'unità, dove  
 $S(k)$  denota il sottoinsieme di  $S$  caratterizzato dai campioni contenenti l'unità in oggetto;  $y_k$ , il valore assunto dalla variabile di interesse  $y$ ;  
 $\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})$  il valore assunto dal vettore di  $J$  variabili ausiliarie.  
 Si vuole stimare il totale  $Y$  della variabile  $y$ , dato dalla seguente espressione:

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

sulla base delle seguenti informazioni:

- per ciascun elemento del campione  $s$  si dispone delle  $J+1$  osservazioni ;

$$(y_k, \mathbf{x}_k)$$

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

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

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

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

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

L'insieme dei pesi finali  $\{w_k; k=1, \dots, n\}$  è ottenuto come soluzione di un problema di minimo vincolato in cui la funzione obiettivo è data da<sup>11</sup>

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

ed i vincoli sono espressi dal sistema di  $J$  equazioni

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

<sup>11</sup> La (4) è equivalente a minimizzare il valore atteso sotto il disegno di campionamento della funzione obiettivo, in quanto la (4) è valida per ciascun campione  $s$  estratto nello spazio campionario.

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

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

- (a) per ogni fissato  $d_k > 0$  esiste un'intervallo  $I(d_k)$ , contenente  $d_k$ , in cui  $G_k(w_k; d_k)$  sia strettamente convessa, differenziabile rispetto a  $w_k$  e non negativa, e si abbia inoltre  $G_k(d_k; d_k) = 0$ ;
- (b) la derivata prima di  $G_k(w_k; d_k)$ , indicata con  $g_k(w_k; d_k) = \frac{\delta G_k(w_k; d_k)}{\delta w_k}$

deve essere una funzione continua e biunivoca nell'intervallo  $I(d_k)$ . Da ciò deriva che  $g_k(w_k; d_k)$  nell'intervallo  $I(d_k)$  è una funzione strettamente crescente di  $w_k$  ed inoltre  $g_k(d_k; d_k) = 0$ . Inoltre, poiché la funzione è strettamente crescente e continua, esiste la funzione inversa,  $g_k^{-1}(\cdot)$ , ovvero una funzione per la quale vale

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

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

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

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

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

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

$$g_k^{-1}(g_k(w_k; d_k)) = g_k^{-1}(x'_k \lambda) \quad ,$$

da cui deriva

$$w_k = g_k^{-1}(x'_k \lambda) \quad (8)$$

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

$$w_k = d_k \gamma_k = g_k^{-1}(x'_k \lambda) \quad ,$$

da cui deriva

$$\gamma_k = \frac{1}{d_k} g_k^{-1}(x'_k \lambda) \quad .$$

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

$$w_k = d_k F_k(x'_k \lambda) \quad (9)$$

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

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

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

esprimibile in termini vettoriali come

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

che è equivalente a

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

Infine, indicando con:

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

è possibile riscrivere il sistema (11) nel seguente modo

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

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

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

Una soluzione numerica<sup>12</sup> al sistema (13) può essere ottenuta in modo iterativo mediante il metodo di Newton. Per illustrare tale metodo, indichiamo con  $v$  ( $v=1, 2, \dots$ ) la generica iterazione e con  $\boldsymbol{\lambda}_v = (\lambda_{1,v}, \dots, \lambda_{j,v}, \dots, \lambda_{J,v})'$  i valori di  $\boldsymbol{\lambda}$  relativi all'iterazione. I passi dell'algoritmo sono i seguenti:

- 1) si pone il valore iniziale di  $\mathbf{l}$ , che indichiamo come  $\mathbf{l}_0$ , pari a  $\boldsymbol{\lambda}_0 = \mathbf{0}$ , dove  $\mathbf{0}$  indica un vettore di dimensione  $J$  i cui elementi sono tutti pari a zero;
- 2) i valori  $\boldsymbol{\lambda}_v$  alle successive iterazioni ( $v = 1, 2, \dots$ ) sono dati da:

$$\boldsymbol{\lambda}_v = \boldsymbol{\lambda}_{v-1} + \left\{ \left[ \frac{\partial \boldsymbol{\phi}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right]_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{v-1}} \right\}^{-1} \{ \mathbf{X} - \tilde{\mathbf{X}} - \boldsymbol{\phi}(\boldsymbol{\lambda}_{v-1}) \}, \quad (15)$$

<sup>12</sup> Altre soluzioni per specifiche funzioni di distanza sono riportate nel lavoro di Singh e Mohl (1996 §)

in cui:

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

- 3) Si itera il passo 2 finchè non viene verificata almeno una delle due condizioni di seguito riportate:

$$\text{Max}_j \left( \frac{|\lambda_{j,v-1} - \lambda_{j,v}|}{|\lambda_{j,v-1}|} \right) \leq \omega \quad (16)$$

$$v = v_{\text{MAX}} , \quad (17)$$

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

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

- 1) per ciascuna unità del campione si calcolano i pesi  $d_k = \pi_k^{-1}$  diretti, ( $k=1,...,n$ ),
- 2) si sceglie la funzione di distanza;  $G_k(w_{ks}; d_k)$

- 3) si definisce la funzione  $g_k(w_{ks}; d_k)$ , ottenuta come derivata prima di  $G_k(w_{ks}; d_k)$ ;
- 4) dall'equazione  $g_k(w_{ks}; d_k) - \mathbf{x}_k' \boldsymbol{\lambda} = 0$  si determina la funzione inversa  $w_{ks} = g_k^{-1}(\mathbf{x}_k' \boldsymbol{\lambda})$ ;
- 5) si ottiene l'espressione funzionale  $F_k(\mathbf{x}_k' \boldsymbol{\lambda}) = \frac{1}{d_k} g_k^{-1}(\mathbf{x}_k' \boldsymbol{\lambda})$  del correttore  $\gamma_{ks}$  del peso diretto;
- 6) si determinano i valori di  $\boldsymbol{\lambda}$ , risolvendo il sistema  $\boldsymbol{\phi}(\boldsymbol{\lambda}) = \mathbf{X} - \sum_{k \in S} d_k \mathbf{x}_k$  secondo il metodo di Newton;
- 7) si calcolano i valori numerici di correttori  $\gamma_{ks}$  ( $k=1, \dots, n$ ) sostituendo i valori di  $\boldsymbol{\lambda}$  nelle espressioni funzionali  $F_k(\mathbf{x}_k' \boldsymbol{\lambda})$ ;
- 8) si determinano i pesi finali mediante il prodotto  $w_{ks} = d_k \gamma_{ks}$ ;
- 9) è possibile, quindi, calcolare lo stimatore di ponderazione vincolata 
$$\tilde{Y}_{PV} = \sum_{k \in S} y_k w_{ks}.$$

### A.1.3. Scelta della funzione di distanza

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

**Tabella 1 - Funzioni di distanza**

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

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

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

- l'utilizzazione della funzione di distanza *lineare* può condurre alla

determinazione di alcuni pesi negativi, in quanto l'intervallo di variazione ammesso per i pesi finali è del tipo  $(-\infty, +\infty)$  ;

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

Al fine di non appesantire la trattazione seguente verranno descritte in dettaglio solamente le funzioni di distanza lineare, logaritmica e logit troncata che si ritengono utili a risolvere la maggior parte dei problemi di stima che si incontrano nelle indagini concrete su larga scala<sup>13</sup> (Singh e Mohl, 1996). Le ragioni di tale scelta risiedono nel fatto che tali funzioni rivestono un'importanza particolare rispetto alle altre; infatti, le funzioni lineare e logaritmica, a differenza delle altre, garantiscono l'ottenimento di una soluzione al sistema di minimo vincolato qualora il sistema dei vincoli sia congruente; inoltre la funzione di distanza logit permette di restringere il campo di variabilità dei pesi finali definendo opportunamente i parametri L ed U.

---

<sup>13</sup> Nei lavori di Deville e Särndal (1992) e di Singh e Mohl (1996 §) vengono introdotte altre funzioni di distanza. In questa sede limitiamo però l'esposizione unicamente i tre metodi adottati nelle indagini Istat ed utili a risolvere la maggior parte dei problemi di stima che si pongono nelle indagini su larga scala



### *Distanza euclidea o lineare*

La funzione di distanza è espressa da

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

in cui  $1/q_k$  indica un peso non correlato a  $d_k$  assegnato all'unità  $k$ -esima. Nella maggior parte delle applicazioni si utilizza il peso uniforme  $1/q_k = 1$ , ma in alcuni casi può essere conveniente utilizzare pesi  $1/q_k$  variabili<sup>14</sup>.

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

$$g_k(w_{ks}; d_k) = -\frac{2}{q_k d_k} (d_k - w_{ks}) \quad (19)$$

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

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

la cui soluzione esplicita è

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

da cui si evince che

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

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

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

---

<sup>14</sup> Nel lavoro di Alexander (1987), si dimostra l'utilità dell'adozione dei pesi  $1/q_k$  variabili, per risolvere particolari problemi di sottocopertura

da cui si ha

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

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

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

che è equivalente a

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

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

### *Distanza logaritmica*

La funzione di distanza è espressa da

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

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

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

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

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

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

la cui soluzione esplicita è

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

da cui si evince che

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

Sostituendo la (24) nella (12) si ottiene

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

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

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

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

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

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

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

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

### *Distanza logit o logaritmica troncata*

La funzione di distanza è espressa da

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

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

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

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

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

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

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

la cui soluzione esplicita rispetto al peso finale è data da

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

da cui si evince che

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

Sostituendo la (29) nella (12) si ottiene

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

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

$$\sum_{k \in S} d_k \mathbf{x}_k \exp(q_k \mathbf{x}'_k \boldsymbol{\lambda}) = \mathbf{X} - \tilde{\mathbf{X}}. \quad (31)$$

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

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

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

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

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

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

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

La (19) è la funzione di distanza che viene utilizzata per l'ottenimento delle stime di massima verosimiglianza dei modelli log-lineari (Darroch and Ratcliff, 1972) ed è stata adottata in ambito ISTAT per il calcolo dei

pesi finali dell'indagine multiscopo sulle Famiglie (ISTAT,1993). Essa porta a coefficienti di correzione che possono variare nell'intervallo  $(0, \infty)$  e conduce quindi a pesi finali sempre positivi. Tuttavia, in alcuni casi non favorevoli, i pesi finali possono presentare valori estremamente grandi rispetto ai corrispondenti pesi base, risultando pertanto non accettabili in quanto la loro applicazione per l'ottenimento di stime riferite a varie sottopopolazioni in differenti domini di studio può condurre a valori non realistici delle stime stesse.

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



## A.2. Stimatore di regressione generalizzata

### A.2.1 Premessa

Il presente capitolo è finalizzato ad illustrare le principali caratteristiche logiche ed algebriche dello stimatore di *regressione generalizzata*, che si fonda sull'utilizzazione di variabili ausiliarie per le quali si conoscono i totali riferiti alla popolazione oggetto d'indagine. Ai fini della costruzione dello stimatore, si adopera l'informazione ausiliaria disponibile ipotizzando un modello di regressione che lega le variabili ausiliarie, che costituiscono le variabili esplicative del modello, alle variabili d'interesse. Una delle caratteristiche più interessanti dello stimatore in oggetto, è quella di poter utilizzare modelli regressivi con caratteristiche differenti; come vedremo meglio in seguito, ciò significa qualificare il modello regressivo in termini dei tre elementi fondamentali di: *gruppo di riferimento del modello*, *livello del modello*, *tipo di modello*<sup>15</sup>.

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

- sono noti i totali della popolazione di una o più variabili ausiliarie;
- viene costruito il modello di regressione sottostante lo stimatore.
- I *gruppi* rappresentano, quindi, una partizione della popolazione di riferimento e per ciascuno di essi si definisce uno specifico modello di regressione.

---

<sup>15</sup> I tre elementi fondamentali che caratterizzano il modello di regressione corrispondono ai concetti di model group, model level e model type, introdotti nell'articolo di Estevao, Hidioglou e Särndal (1995 §)



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

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

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

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

Per illustrare le caratteristiche degli stimatori di regressione generalizzata, introduciamo livelli crescenti di complessità: a tal fine, nel successivo *paragrafo* 4.2.2 viene introdotto lo stimatore di regressione generalizzata con

riferimento al caso più semplice di modello definito al livello di elemento e di un unico gruppo di riferimento del modello, costituito da tutta la popolazione. Successivamente, nel *paragrafo A.2.3*, dopo aver approfondito il concetto di *livello del modello*, viene introdotto il modello di regressione al livello di cluster. Nel *paragrafo A.2.4* è sviluppato il tema del *gruppo di riferimento del modello*; infine, nel *paragrafo A.2.5* viene svolta una trattazione più approfondita del concetto di *tipo di modello*.

## A.2.2 Modello a livello di unità elementari

### A.2.2.1 Simbologia e prima formulazione dello stimatore di regressione generalizzata

Sia  $U$  una popolazione finita di  $N$  unità elementari, che indichiamo come  $U = \{1, \dots, k, \dots, N\}$ , e sia  $s$  un campione casuale di  $n$  elementi, che indichiamo come  $s = \{1, \dots, k, \dots, n\}$ , estratto da  $U$  mediante un disegno di campionamento che genera l'*universo dei campioni*  $S$  (ossia l'insieme di tutti i possibili campioni estraibili mediante il disegno in parola) ed assegna al generico campione  $s$  la probabilità  $p(s)$  di essere estratto, dove  $\sum_{s \in S} p(s) = 1$ .

Con riferimento alla generica unità  $k \in U$ , indichiamo quindi con:  $\pi_k = \sum_{s \in S(k)} p(s)$ , la probabilità di inclusione nel campione dell'unità, dove

$S(k)$  denota il sottoinsieme di  $S$  caratterizzato dai campioni contenenti l'unità in oggetto;  $y_k$ , il valore assunto dalla variabile di interesse  $y$ ;  $x_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})$ , il valore assunto dal vettore di  $J$  variabili ausiliarie. Si vuole stimare il totale  $Y$  della variabile  $y$ , dato dalla seguente espressione:

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

sulla base delle seguenti informazioni:

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

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

E' utile chiarire che le  $J$  variabili ausiliarie vengono individuate cercando, nell'insieme delle variabili per le quali sono noti i totali al livello di popolazione, le variabili maggiormente correlate con la variabile  $y$  di interesse; in tal modo, il vettore di variabili ausiliarie fornisce informazioni sulla variabile  $y$  di cui si può tenere conto nella fase di costruzione dello stimatore. Introduciamo, quindi, un modello di regressione lineare, che indichiamo con  $\xi$ , per spiegare la forma della nuvola di punti definita sugli  $N$  elementi della popolazione finita  $U$

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

Il modello si basa sulle seguenti assunzioni:

- (i) i valori  $y_1, \dots, y_k, \dots, y_N$  assunti dalla variabile  $y$  per le  $N$  unità della popolazione sono considerati come realizzazioni di  $N$  variabili casuali indipendenti;
- (ii) le variabili ausiliarie sono trattate come costanti note di tipo non stocastico;
- (iii) la relazione che lega la generica variabile casuale  $y_k$  con il vettore  $\mathbf{x}_k$  ( $k=1, \dots, N$ ) è la seguente

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

in cui

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_J)'$$

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

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

essendo  $c_k$  (per  $k \in U$ ) delle costanti note.

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

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

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

Ciò premesso, si supponga di aver effettuato un censimento di tutte le  $N$  unità della popolazione  $U$  e di disporre, quindi, di tutti i valori della nuvola di punti (3), si supponga, inoltre, che la nuvola di punti osservata si adatti piuttosto bene al modello appena introdotto. E' possibile utilizzare, allora, la nuvola di punti della popolazione per stimare, mediante il metodo dei minimi quadrati ponderati il vettore dei coefficienti di regressione  $\beta$  del modello  $\xi$ . Utilizzando la teoria standard della regressione generalizzata, si ha che il miglior stimatore lineare non distorto dei coefficienti  $\beta$ , sotto il modello  $\xi$ , è dato da

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

Il vettore<sup>16</sup> dei coefficienti  $\mathbf{B}$  è, ovviamente, una caratteristica incognita della popolazione. E' possibile, tuttavia, stimare  $\mathbf{B}$ , mediante i dati rilevati dal campione  $s$ . La relazione (7) si presenta come il prodotto di totali della popolazione ed una sua stima asintoticamente corretta può essere ottenuta, stimando correttamente ciascun totale mediante lo stimatore di Horvitz-Thompson. Siano, infatti

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

---

<sup>16</sup> La stima (7) è derivata mediante il metodo dei minimi quadrati, che, come è noto porta a definire il migliore stimatore lineare corretto

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

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

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

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

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

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

In sintesi, quindi, una stima asintoticamente corretta<sup>17</sup> della (7) è data da:

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

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

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

$$\{(y_k, x_{1k}, \dots, x_{jk}, \dots, x_{Jk}) : k = 1, \dots, s\} \quad (13)$$

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

- a) con riferimento alle  $N$  unità della popolazione, i valori interpolati,

<sup>17</sup> La stima è solo asintoticamente corretta in quanto il valore atteso dell'inversa di una matrice ad elementi casuali è diverso dall'inversa del valore atteso della matrice stessa. In formule ciò è espresso da

$$E(\tilde{\mathbf{T}}_1^{-1}) = \mathbf{T}_1^{-1} \quad , \quad E(\tilde{\mathbf{T}}_1^{-1}) \neq \tilde{\mathbf{T}}_1^{-1}$$

$\tilde{y}_1, \dots, \tilde{y}_k, \dots, \tilde{y}_N$  relativi ai corrispondenti valori  $y_1, \dots, y_k, \dots, y_N$ ,  
mediante la relazione

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

b) con riferimento alle  $n$  unità del campione i residui

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

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

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

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

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

Dalla (17) risulta che lo stimatore di regressione generalizzata può essere espresso mediante la somma di due totali. Il primo è la somma degli  $N$  valori della popolazione stimati in base alla relazione (14) e contiene l'informazione ausiliaria disponibile al livello dei totali noti,  $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)'$ , della popolazione. Il totale  $\sum_{k \in U} \tilde{y}_k$  può essere, infatti, riscritto mediante il seguente passaggio:

$$\sum_{k \in U} \tilde{y}_k = \sum_{k \in U} \mathbf{x}'_k \tilde{\mathbf{B}} = \left( \sum_{k \in U} \mathbf{x}_k \right)' \tilde{\mathbf{B}} = \mathbf{X}' \tilde{\mathbf{B}} \quad (18)$$

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

$$\begin{aligned}\sum_{k \in s} \tilde{e}_k &= \sum_{k \in s} \frac{(y_k - \mathbf{x}_k' \tilde{\mathbf{B}})}{\pi_k} = \\ &= \sum_{k \in s} \left( \frac{y_k}{\pi_k} \right) - \sum_{k \in s} \left( \frac{\mathbf{x}_k}{\pi_k} \right)' \tilde{\mathbf{B}} = \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}' \tilde{\mathbf{B}},\end{aligned}\quad (19)$$

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

$$\tilde{\mathbf{Y}} = \sum_{k \in s} \left( \frac{y_k}{\pi_k} \right), \quad \tilde{\mathbf{X}} = \sum_{k \in s} \left( \frac{\mathbf{x}_k}{\pi_k} \right)$$

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

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

dalla quale risulta che tale stimatore è ottenuto come somma dello stimatore di Horvitz-Thompson del totale  $\mathbf{Y}$  più un termine di aggiustamento regressivo che dipende dalle differenze tra totali noti le corrispondenti stime campionarie di Horvitz-Thompson, ponderate con i rispettivi coefficienti di regressione stimati. Per calcolare lo stimatore di regressione generalizzata in base all'espressione (20) è necessario conoscere i totali delle variabili ausiliarie  $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)'$  ed i valori della variabile di interesse e delle variabili ausiliarie  $\{(y_k, x_{1k}, \dots, x_{jk}, \dots, x_{Jk}) : k = 1, \dots, s\}$  per le  $n$  unità campionate. Pertanto, non è necessario conoscere i valori delle variabili ausiliarie per le unità della popolazione non campionate<sup>18</sup>.

<sup>18</sup> Questo è il caso delle indagini in cui sono disponibili unicamente i totali delle variabili ausiliarie mentre i singoli valori vengono rilevati con l'indagine campionaria. In tale caso, ovviamente, le variabili mediante le quali vengono costruiti i totali noti e le variabili rilevate sulle singole unità devono avere la medesima definizione concettuale ed essere riferite allo stesso periodo temporale. E' chiaro che l'allontanamento da tale condizione introduce nella stima fattori distorsivi

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

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

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

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

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

#### **A.2.2.2 Espressioni alternative dello stimatore di regressione generalizzata**

*Espressione in termini dei pesi*

Un'espressione alternativa dello stimatore di regressione generalizzata è data da

$$\tilde{Y}_{\text{REG}} = \sum_{k \in S} y_k d_k \gamma_k = \sum_{k \in S} y_k w_k, \quad (22)$$

in cui si è denotato con:

$$w_k = d_k \gamma_k, \quad \text{il peso finale,}$$

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

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

La (22) permette di esprimere lo stimatore di regressione generalizzata come una somma ponderata, con i pesi  $d_k \gamma_k$  (detti *pesi finali*), dei dati campionari  $y_1, \dots, y_k, \dots, y_n$ . Tale espressione è formalmente simile a quella dello stimatore di Horvitz - Thompson che, come è noto, è espresso come somma pesata, con i pesi base, dei dati campionari. Occorre, tutta-



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

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

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

in cui i vincoli sono dati da:

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

Da tale sistema si ottiene

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

*Espressione utile per il calcolo della varianza*

Le espressioni alternative (17), (20) e (22) dello stimatore di regressione generalizzata sopra introdotte permettono di illustrare differenti aspetti e

caratteristiche dello stimatore stesso. Tuttavia è importante aggiungere ad esse un'ultima espressione che sarà utile per derivare agevolmente la formula della varianza dello stimatore. A tal fine, si introducono le seguenti quantità

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

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

$$\tilde{Y}_{\text{REG}} = \sum_{k \in S} \frac{\gamma_k (y_k^* + e_k^*)}{\pi_k} = \sum_{k \in S} \frac{\gamma_k y_k^*}{\pi_k} + \sum_{k \in S} \frac{\gamma_k e_k^*}{\pi_k} \quad (25)$$

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

$$\left. \sum_{\epsilon} \frac{\gamma_{\epsilon} y_{\epsilon}^*}{\pi_{\epsilon}} \right| = \sum_{\epsilon} \frac{\gamma_{\epsilon} y_{\epsilon}^*}{\pi_{\epsilon}} = \sum_{\epsilon} \frac{\gamma_{\epsilon} y_{\epsilon}^*}{\pi_{\epsilon}} \quad (26)$$

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

$$\tilde{Y}_{\text{REG}} = \sum_{\epsilon} \frac{\gamma_{\epsilon} y_{\epsilon}^*}{\pi_{\epsilon}} + \sum_{k \in S} \frac{\gamma_k e_k^*}{\pi_k} \quad (27)$$

### A.2.2.3 Alcune considerazioni sul ruolo del modello

È importante svolgere alcune considerazioni circa il ruolo del modello  $\xi$  nella costruzione dello stimatore di regressione generalizzata. Non è possibile, tuttavia, dare una dimostrazione formale delle proprietà enunciate poiché non è stata introdotta, ancora, la formula della varianza dello stimatore. Lo stimatore è distorto ma è asintoticamente corretto, e, quindi, per campioni sufficientemente grandi (come quelli che caratterizzano le indagini effettuate dall'Istituto Nazionale di Statistica) si può assumere che lo stimatore sia corretto. Inoltre lo stimatore è consistente nel senso che la sua varianza tende ad annullarsi al crescere della dimensione campionaria.

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

L'efficienza dello stimatore in parola, in confronto a quella dello stimatore di Horvitz-Thompson, è funzione inversa dei residui  $e_k^*$ , e quindi dipende dal grado di adattamento della nuvola di punti della popolazione alla retta di regressione. La proprietà di consistenza sotto il disegno delle stime ottenute mediante lo stimatore di regressione generalizzata e la validità della formula della varianza non dipendono, tuttavia, dal fatto se il modello sia valido oppure no. Da ciò deriva, in particolare, che l'inferenza prodotta è *assistita* dall'introduzione di un modello ma non è *dipendente* da esso in quanto tale (nella letteratura in lingua anglosassone, con riferimento ai due concetti appena introdotti, si usano rispettivamente i termini: *model assisted* e *model based*).

## A.2.3 Livello del modello

### A.2.3.1 Introduzione al problema

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

ausiliarie si riferiscono a grappoli di elementi ed i totali si ottengono come somma, sulla popolazione dei grappoli, delle variabili ausiliarie relative ai grappoli. Per meglio illustrare tale aspetto consideriamo gli esempi di seguito riportati.

### *Esempio 1*

La popolazione oggetto di indagine è costituita dagli individui; si effettua un piano di campionamento a due stadi in cui si selezionano al primo stadio i comuni e al secondo stadio le famiglie e si osservano le variabili oggetto di indagine su tutti gli individui appartenenti alle famiglie campionate. Le variabili ausiliarie, riferite agli individui, sono il *sex* e l'*età*; i totali noti sono definiti dalla distribuzione della popolazione per sesso ed età. Nella situazione appena descritta gli individui sono le *unità elementari* e le famiglie costituiscono *grappoli di unità*.

### *Esempio 2*

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

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

- a) le informazioni ausiliarie sono disponibili a livello di elemento (vedi esempio 1);
- b) le informazioni ausiliarie sono disponibili solamente al livello di grappolo mentre non sono note tali informazioni per ciascun elemento appartenente al grappolo (vedi esempio 2).

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

Nel caso in cui si adottino disegni a più stadi di campionamento (come nell'esempio 1), a seconda della disponibilità dell'informazione ausiliaria, è possibile definire il livello del modello in modo differente: ad esempio è possibile individuare:

- (i) un modello a livello di elemento;
- (ii) un modello a livello di unità primaria;
- (iii) un modello a livello di grappoli di unità elementari selezionati all'ultimo stadio di campionamento;
- (iv) un modello a più livelli di riferimento. Ad esempio a livello di elemento e a livello di unità primaria.

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

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

### **A.2.3.2 Campionamento a grappoli**

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

- (i) si seleziona un campione  $s_I = \{1, \dots, i, \dots, n_I\}$  di  $n_I$  grappoli mediante il disegno di campionamento che genera l'universo dei campioni  $S_I$  e assegna al generico campione la probabilità  $p_I(s_I)$  di essere estratto (dove  $\sum_{s_I \in S_I} p_I(s_I) = 1$ );
- (ii) di conseguenza, indicando con  $S_I(i)$  il sottoinsieme di  $S_I$  formato dai campioni contenenti il grappolo  $i$ -esimo la probabilità d'inclusione di tale grappolo è data da  $\pi_{II} = \sum_{s_I \in S_I(i)} p_I(s_I)$ ;
- (iii) tutte le unità elementari dei grappoli selezionati vengono incluse nel campione; tale circostanza determina il fatto che la probabili-

tà d'inclusione delle unità elementari coincide con quella dei grappoli di appartenenza; pertanto, la probabilità di inclusione  $\pi_{lik}$  dell'unità elementare  $k$  appartenente al grappolo  $i$  è data da  $\pi_{lik} = \pi_{li}$  (per  $k=1, \dots, N_{Ii}$ , e per  $i=1, \dots, N_I$ ).

Facendo riferimento al  $k$ -esimo elemento del grappolo  $i$  (per  $k=1, \dots, N_{Ii}$ , e per  $i=1, \dots, N_I$ ) indichiamo quindi con  $y_{lik}$  il valore della variabile d'interesse  $y$  e con  $\mathbf{x}_{lik}$  il valore assunto dal vettore di  $J$  variabili ausiliarie; considerando, quindi, il grappolo<sup>19</sup> nel suo complesso si ha:

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

### *Modello a livello di grappolo*

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

$$Y = \sum_{i=1}^{N_I} \sum_{k=1}^{N_{Ii}} y_{lik} = \sum_{i=1}^{N_I} y_{li} \quad , \quad (28)$$

utilizziamo il seguente modello  $\xi$  a livello di grappolo

$$y_{li} = \mathbf{x}_{li}' \boldsymbol{\beta}_I + \varepsilon_{li} \quad \text{per } (i=1, \dots, N_I) \quad (29)$$

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

$$E_{\xi}(\varepsilon_{li}) = 0 \quad , \quad \text{Var}_{\xi}(\varepsilon_{li}) = c_{li} \sigma^2 \quad , \quad \text{Cov}_{\xi}(\varepsilon_{li}, \varepsilon_{li'}) = 0 \quad \text{per } \forall i \neq i' \quad (30)$$

essendo  $c_{li}$  (per  $i=1, \dots, N_I$ ) delle costanti note.

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

<sup>19</sup> E' chiaro che nella situazione b), illustrata nel precedente paragrafo A.2.3.1, è possibile definire unicamente il valore delle variabili ausiliarie a livello di grappolo  $\mathbf{x}_{li}$  e non il valore  $\mathbf{x}_{lik}$  per ciascuno degli elementi del grappolo

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

in cui si è denotato con:

$$\begin{aligned}
w_{Ii} &= d_{Ii} \gamma_{Ii} \quad , & \text{il peso finale,} \\
d_{Ii} &= \frac{1}{\pi_{Ii}} \quad , & \text{il peso diretto,} \\
\gamma_{Ii} &= 1 + (\mathbf{X}_I - \tilde{\mathbf{X}}_I)' \left( \sum_{i=1}^{n_I} \frac{d_{Ii} \mathbf{x}_{Ii} \mathbf{x}_{Ii}'}{c_{Ii}} \right)^{-1} \frac{\mathbf{x}_{Ii}}{c_{Ii}} & \text{il fattore correttivo del peso base,}
\end{aligned}$$

essendo

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

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

#### *Modello a livello di unità elementare*

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

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

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

$$E_{\xi}(\varepsilon_{lik}) = 0, \text{ Var}_{\xi}(\varepsilon_{lik}) = c_{lik} \sigma^2, \text{ Cov}_{\xi}(\varepsilon_{lik}, \varepsilon_{li'k'}) = 0 \quad \text{per } \forall ik \neq i'k', \quad (33)$$

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

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

in cui si è denotato con

$$\begin{aligned} w_{lik} &= d_{li} \gamma_{lik}, & \text{il peso finale,} \\ d_{li} &= \frac{1}{\pi_{li}}, & \text{il peso diretto,} \\ \gamma_{lik} &= 1 + (\mathbf{X} - \tilde{\mathbf{X}})' \left( \sum_{i=1}^{n_I} \sum_{k=1}^{N_{li}} \frac{d_{li} \mathbf{x}_{lik} \mathbf{x}_{lik}'}{c_{lik}} \right)^{-1} \frac{\mathbf{x}_{lik}}{c_{lik}} & \text{il fattore correttivo del peso base} \end{aligned}$$

dove

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

La definizione di un modello a livello di unità elementare comporta, quindi, il fatto che a ciascuna unità di un grappolo venga attribuito un peso finale differente<sup>20</sup>.

### *Scelta del livello del modello*

Un campionamento a grappoli consente di stimare parametri riferiti sia alla popolazione delle unità elementari che alla popolazione dei grappoli. Per illustrare tale aspetto consideriamo, ad esempio, un'indagine sulle famiglie in cui le famiglie costituiscono i grappoli e gli individui le unità elementari e supponiamo di avere rilevato per ciascuna famiglia una variabile dicotomica  $y_{li}$  che assume valore 1 nel caso che la famiglia abbia tre componenti e valore 0 altrimenti; supponiamo, inoltre, di avere rilevato

<sup>20</sup> Questo non è vero nel caso in cui tutte le unità del grappolo presentino lo stesso valore delle variabili ausiliarie



per ogni individuo una variabile dicotomica  $y_{lik}$  che assume valore 1 nel caso in cui esso viva in una famiglia di tre componenti e valore 0 altrimenti; utilizzando la variabile  $y_{li}$  è possibile ottenere una stima del numero di famiglie con tre componenti, mentre utilizzando le variabili  $y_{lik}$  si può calcolare la stima del numero di persone che vivono in famiglie di tre componenti; è ovvio che questa ultima stima divisa per tre fornisce, nuovamente, una stima del numero di famiglie di tre componenti. L'esempio appena introdotto, mostra un caso molto frequente nelle indagini ISTAT sulle famiglie in cui è possibile ottenere una stima di uno stesso parametro oggetto di indagine (ad esempio, il numero di famiglie di tre componenti) sia utilizzando le informazioni relative agli individui sia quelle relative alle famiglie; nasce da qui l'esigenza di *coerenza* tra l'insieme delle stime riferite alle unità elementari e quelle riferite ai grappoli. Tale coerenza, implica quindi l'uguaglianza delle stime relative allo stesso parametro incognito di popolazione e si ottiene attribuendo il medesimo peso finale al grappolo ed a tutte le unità elementari ad esso appartenenti.

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

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

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

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

ri appartenenti al grappolo; risulta quindi difficile attribuire un peso finale al grappolo differente dal suo peso diretto<sup>21</sup>.

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

---

**Sintetizziamo nella seguente tabella i criteri di scelta appena descritti**

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

---

#### **A.2.3.3 Disegni di campionamento a due o più stadi**

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

Ipotizziamo, inoltre, che la  $i$ -esima UP sia costituita da  $N_{II}$  *Unità Secondarie* (US). Le US possono essere unità elementari o alternativamente grappoli di unità elementari. Indicando quindi con  $y_{Iik}$  il valore della variabile

---

<sup>21</sup> Questo problema può essere risolto, anche, mediante il metodo noto nelle letterature in lingua anglosassone con il termine *principal person method* che assegna al grappolo il peso finale calcolato per l'unità elementare più rappresentativa o più importante del grappolo Alexander (1987)

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

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

essendo

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

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

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

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

con  $S_I(i)$  il sottoinsieme di  $S_I$  formato dai campioni contenenti la UP  $i$ -esima, la probabilità d'inclusione di tale UP è data da

$$\pi_{Ii} = \sum_{s_I \in S_I(i)} p_I(s_I);$$

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

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

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

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

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

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

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

$$E_{\xi}(\varepsilon_{Ii}) = 0, \quad \text{Var}_{\xi}(\varepsilon_{Ii}) = c_{Ii} \sigma^2, \quad \text{Cov}_{\xi}(\varepsilon_{Ii}, \varepsilon_{Ii'}) = 0 \quad \text{per } \forall i \neq i', \quad (37)$$

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

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

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

in cui si è denotato con:

$$w_{lik} = d_{lik} \gamma_{li} \quad , \quad \text{il peso finale,}$$

$$d_{lik} = \frac{1}{\pi_{lik}} \quad , \quad \text{il peso diretto,}$$

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

essendo

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

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

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

$$\tilde{\mathbf{x}}_{li} = \sum_{i=1}^{n_{li}} \frac{\mathbf{x}_{lik}}{\pi_{lik|li}} \quad .$$

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

strategia campionaria in cui le US siano selezionate nel secondo stadio con probabilità uguale ed in cui si adotti un modello a livello di UP.

#### A.2.4. Gruppo di riferimento del modello

##### A.2.4.1 Modello a livello di unità elementare

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

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

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

$$U = \bigcup_{g=1}^G U_{(g)} \quad \text{e} \quad \emptyset = U_{(g)} \cap U_{(g')} \quad (\text{per } g \neq g' = 1, \dots, G)$$

da cui deriva

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

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

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

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

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

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

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

$$E_\xi(\varepsilon_k) = 0, \quad \text{Var}_\xi(\varepsilon_k) = c_k \sigma^2, \quad \text{Cov}_\xi(\varepsilon_k, \varepsilon_{k'}) = 0 \quad \text{per } \forall k \neq k'; \quad (40)$$

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

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

in cui si è indicato con:

$$w_k = d_k \gamma_k, \quad \text{il peso finale,}$$

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

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

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

essendo

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

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

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

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

$$\mathbf{z}_k' = (\delta_{(1)k} \mathbf{x}_k', \dots, \delta_{(g)k} \mathbf{x}_k', \dots, \delta_{(G)k} \mathbf{x}_k') \quad (43)$$

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

dove  $\delta_{(g)k}$  ( $g = 1, \dots, G$ ) è una variabile dicotomica che assume valore 1 se l'unità  $k$ -esima appartiene al gruppo  $g$  e valore 0 altrimenti e i vettori  $\mathbf{z}_k'$  e  $\boldsymbol{\beta}$  sono costituiti da  $A = J \times G$  elementi. Il vettore  $\boldsymbol{\delta}_k = (\delta_{(1)k}, \dots, \delta_{(g)k}, \dots, \delta_{(G)k})'$  che contiene le variabili indicatrici  $\delta_{(g)k}$  appena introdotte ha  $J-1$  termini pari a zero ed un singolo termine pari ad 1, che identifica il gruppo al quale il  $k$ -esimo l'elemento appartiene; è valida, pertanto, la seguente relazione

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

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

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

$$\tilde{Y}_{\text{REG}} = \sum_{k=1}^n y_k d_k \tilde{a}_k = \sum_{k=1}^n y_k w_k \quad (45)$$

in cui si è denotato con:

$$w_k = d_k \tilde{a}_k, \quad \text{il peso finale,}$$

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

dove

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



essendo la matrice

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

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

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

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

$$\begin{aligned} \gamma_k &= 1 + [\mathbf{x}'_{(1)} - \tilde{\mathbf{x}}'_{(1)}, \dots, \mathbf{x}'_{(g)} - \tilde{\mathbf{x}}'_{(g)}, \dots, \mathbf{x}'_{(G)} - \tilde{\mathbf{x}}'_{(G)}] \begin{bmatrix} \mathbf{Q}_{(1)} & 0 & \dots & 0 \\ 0 & & & \vdots \\ \vdots & 0 & \mathbf{Q}_{(g)} & 0 \\ \vdots & & & 0 \\ 0 & \dots & \dots & 0 & \mathbf{Q}_{(G)} \end{bmatrix} \frac{1}{c_k} \begin{bmatrix} 0 \\ \vdots \\ \mathbf{x}_k \\ \vdots \\ 0 \end{bmatrix} \\ &= 1 + (\mathbf{X}_{(g)} - \tilde{\mathbf{X}}_{(g)})' \mathbf{Q}_{(g)} \frac{\mathbf{x}_k}{c_k} \\ &= 1 + (\mathbf{X}_{(g)} - \tilde{\mathbf{X}}_{(g)})' \left( \sum_{k=1}^{n(g)} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} \quad (\text{per } k \in s_{(g)}). \end{aligned}$$

I post-strati definiscono un importante caso di gruppi di riferimento del modello poiché, essi, per definizione costituiscono delle sub-popolazioni non sovrapposte per le quali sono noti i totali di riferimento; altre sub-popolazioni spesso considerate nel formare i gruppi possono essere gli strati, o aggregazioni di strati costituenti dei domini di stima. Ovviamente affinché queste sub-popolazioni possano essere qualificate come gruppi è necessario che siano noti i totali di riferimento a livello di ciascuna sottopopolazione.

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

- i gruppi costituiscono domini d'interesse, per cui si desidera che le stime a livello di gruppo dei totali di alcune variabili ausiliarie (che

costituiscono le variabili strutturali della popolazione) coincidano con i totali noti;

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

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

con

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

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

### *Esempio*

Si consideri una popolazione di individui raggruppata in  $G$  gruppi  $U(1), \dots, U(g), \dots, U(G)$  contenenti rispettivamente  $N(1), \dots, N(g), \dots, N(G)$  individui, dove i gruppi sono definiti in base alle modalità incrociate del sesso e delle classi di età. Si supponga, inoltre, di disporre, per ciascuna unità  $k$ , di una variabile ausiliaria  $x_k$  di cui sono noti i valori del totale per ciascun gruppo

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

Sotto il modello

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

con

$$E_{\xi}(\varepsilon_k) = 0 \quad ; \quad \text{Var}_{\xi}(\varepsilon_k) = x_k \sigma^2 \quad ; \quad \text{Cov}_{\xi}(\varepsilon_k, \varepsilon_{k'}) = 0 \quad \text{per } k \neq k' ,$$

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

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

a livello di gruppo è uguale a

$$\begin{aligned} \tilde{Y}_{(g)REG} &= \sum_{k \in s} y_k d_k \gamma_k \\ &= \sum_{k=1}^{n_{(g)}} y_k d_k \left( 1 + (X_{(g)} - \sum_{k=1}^{n_{(g)}} d_k x_k) \left( \sum_{k=1}^{n_{(g)}} \frac{d_k x_k^2}{x_k} \right)^{-1} \frac{x_k}{x_k} \right) \\ &= \sum_{k=1}^{n_{(g)}} y_k d_k \left( 1 + (X_{(g)} - \tilde{X}_{(g)}) \tilde{X}_{(g)}^{-1} \right) \\ &= \sum_{k=1}^{n_{(g)}} y_k d_k \frac{X_{(g)}}{\tilde{X}_{(g)}} . \end{aligned}$$

Lo stimatore del totale  $Y$  è, pertanto, dato da

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

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

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

dove

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

#### A.2.4.2 Modello a livello di grappolo

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

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

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

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

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

$S_I(i)$  il sottoinsieme di  $S_I$  formato dai campioni contenenti il grappolo  $i$ -esimo, la probabilità d'inclusione di tale grappolo è data da;

$$\pi_{li} = \sum_{s_I \in S(i)} p_I(s_I) ;$$

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

Supponiamo, infine, che il campione  $s_{I(g)}$  del gruppo  $g$ -definito

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

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

definendo un modello separato per i grappoli di ciascun gruppo:

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

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

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

essendo  $c_{li}$  delle costanti note.

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

$$\begin{aligned} \tilde{Y}_{\text{REG}} &= \sum_{g=1}^G \sum_{i=1}^{N_{I(g)}} \gamma_{li} d_{li} y_{li} = \sum_{g=1}^G \sum_{i=1}^{N_{I(g)}} w_{li} y_{li} \\ &= \sum_{g=1}^G \sum_{i=1}^{N_{I(g)}} \sum_{k=1}^{N_{Ii}} \gamma_{li} d_{li} y_{lik} = \sum_{g=1}^G \sum_{i=1}^{N_{I(g)}} \sum_{k=1}^{N_{Ii}} w_{li} y_{lik} \end{aligned} \quad (48)$$

in cui si è indicato con:

$$w_{li} = d_{li} \gamma_{li}, \quad \text{il peso finale,}$$

$$d_{li} = \frac{1}{\pi_{li}}, \quad \text{il peso diretto,}$$

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

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

essendo

$$\tilde{\mathbf{X}}_{I(g)} = \sum_{i=1}^{nI(g)} \mathbf{x}_{li} d_{li}$$

### A.2.5. Tipo di modello

La definizione del *tipo di modello* consiste nella individuazione del modello di regressione, scegliendo in modo opportuno le variabili ausiliarie e le costanti ( $c_k$  per i modelli a livello di elemento o  $c_{li}$  per i modelli a livello di grappolo) che specificano la variabilità dei residui. Dalla definizione congiunta del *tipo*, del *gruppo* e del *livello del modello* è possibile fare discendere i più importanti stimatori utilizzati nelle indagini campionarie su larga scala. Per illustrare questo aspetto, nel presente paragrafo prenderemo in esame a scopo didattico gli stimatori: *diretto*, *rapporto semplice*, *rapporto post-stratificato*, *ratio-raking* che adottano un modello a livello di unità elementare e che possono essere ottenuti come caso particolare a partire dall'espressione generale dello stimatore di regressione:

$$\tilde{Y}_{\text{REG}} = \sum_{k \in S} y_k d_k \gamma_k, \quad (49)$$

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

#### *Stimatore diretto*

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

- 1) per la generica unità  $k$ -esima il vettore delle variabili ausiliarie contiene un solo elemento che assume valore uguale alla probabilità d'inclusione  $\pi_k$ , inoltre la costante  $c_k$  è uguale a  $\pi_k$ ;
- 2) il vettore dei totali noti delle variabili ausiliarie è costituito da un solo elemento ed è dato da  $\mathbf{X} = \sum_{k \in U} \pi_k = n$

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

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

$$\begin{aligned}\gamma_k &= 1 + (n - \sum_{k \in s} d_k \pi_k) \left( \sum_{k \in s} d_k \pi_k \right)^{-1} \frac{\pi_k}{\pi_k} \\ &= 1 + \frac{(n - n)}{n} = 1\end{aligned}\quad (50)$$

Sostituendo<sup>22</sup>, infine, l'espressione di  $\gamma_k$ , appena ottenuta, nella (49) si ottiene la ben nota espressione dello stimatore diretto

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

#### *Stimatore rapporto semplice*

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

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

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

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

---

<sup>22</sup> È chiaro che i fattori correttivi sono pari a 1 solo nel caso in cui tutte le  $n$  unità del campione sono rispondenti all'indagine. Invece, nel caso in cui sono rispondenti all'indagine solamente  $n_e < n$  unità campionarie per ottenere lo stimatore diretto occorre utilizzare il seguente tipo di modello alternativo  $x_k = c_k = pk(n/n_e)$

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

$$\begin{aligned}
 \gamma_k &= 1 + (X - \tilde{X}) \left( \sum_{k \in s} \frac{d_k x_k^2}{x_k} \right)^{-1} \frac{x_k}{x_k} \\
 &= 1 + (X - \tilde{X}) \left( \sum_{k \in s} d_k x_k \right)^{-1} \\
 &= 1 + \frac{(X - \tilde{X})}{\tilde{X}} = \frac{X}{\tilde{X}} .
 \end{aligned} \tag{52}$$

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

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

#### *Stimatore rapporto post-stratificato*

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

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

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



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

$$\begin{aligned} \gamma_k &= 1 + (X_{(g)} - \sum_{k=1}^{n_{(g)}} d_k x_k) \left( \sum_{k=1}^{n_{(g)}} \frac{d_k x_k^2}{x_k} \right)^{-1} \frac{x_k}{x_k} \\ &= \left( 1 + (X_{(g)} - \tilde{X}_{(g)}) \tilde{X}_{(g)}^{-1} \right) \\ &= \frac{X_{(g)}}{\tilde{X}_{(g)}} \quad \text{per } (k \in s_{(g)}). \end{aligned} \quad (54)$$

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

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

in cui

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

Facciamo notare che lo stimatore<sup>23</sup> (55) definisce come caso particolare tutta una serie di stimatori ben noti nella letteratura sul campionamento, infatti a seconda di come vengono formati i gruppi si hanno:

- *lo stimatore rapporto semplice*, nel caso in cui tutta la popolazione definisca un unico gruppo;
- *lo stimatore del rapporto separato*, nel caso in cui ciascun gruppo sia costituito da un unico strato;
- *lo stimatore del rapporto combinato* nel caso in cui i gruppi siano costruiti come aggregazione di strati;

---

<sup>23</sup> È utile osservare che lo stimatore espresso dalla (55) può essere ottenuto in modo alternativo a quanto appena fatto utilizzando il modello definito dalle espressioni (41)-(43) in cui tutta la popolazione costituisce un unico gruppo

- lo stimatore del rapporto *post-stratificato*, nel caso in cui le  $G$  sottopopolazioni che costituiscono i gruppi siano *post-strati*. Si assume, in tal caso, che la variabile utilizzata per definire la partizione in gruppi non sia stata usata per la stratificazione delle unità, ma venga rilevata per ciascuna unità elementare inclusa nel campione; ciò implica, in particolare, che il numero di unità campionarie ricadenti in ciascun *post-strato* è una variabile casuale e ciascun *post-strato* è costituito dall'unione di parti di strati del disegno di campionamento.

### *Stimatore ratio-raking*

Definiamo questo tipo di stimatore prendendo in esame un modello a livello di unità elementari ed esaminiamo la situazione di una popolazione suddivisa in  $G$  gruppi in cui, per il generico gruppo  $g$  ( $g=1,...,G$ ), si possano individuare due partizioni distinte. La prima partizione composta di  $R$  sottopopolazioni definite sulla base delle modalità assunte dalla variabile  $x_1$ , mentre la seconda partizione è composta di  $C$  sottopopolazioni definita sulla base delle modalità assunte dalla variabile  $x_2$ . Per ciascun gruppo  $g$  ( $g=1,...,G$ ), il numero di elementi della popolazione appartenenti alla sottopopolazione  $r$  ( $r=1,...,R$ ) della prima partizione è indicato con  $N_{(g),1r}$ ; mentre, si denota con  $N_{(g),2c}$  il numero di elementi della popolazione appartenenti alla sottopopolazione  $c$  ( $c=1,...,C$ ) della seconda partizione; supponiamo inoltre che le quantità  $N_{(g),1r}$  e  $N_{(g),2c}$  siano note.

I dati del problema possono essere riassunti nel modo seguente:

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

$$X'_d = (N_{d,(1,1)}, ..., N_{d,(1,r)}, ..., N_{d,(1,R)}, N_{d,(2,1)}, ..., N_{d,(2,c)}, N_{d,(2,C)})'$$

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

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

dove  $\delta_{k,1r}$  è una variabile indicatrice che assume valore 1 se l'unità  $k$ -esima appartiene alla  $r$ -esima sottopopolazione della prima partizione e

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

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

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

$$\times \begin{bmatrix} \mathbf{A}_{RR} & \mathbf{A}_{RC} \\ \mathbf{A}'_{RC} & \mathbf{A}_{CC} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_{r-1} \\ 1 \\ \mathbf{0}_{R+c-r-1} \\ 1 \\ \mathbf{0}_{C-c} \end{bmatrix} \quad (\text{per } (k \in s_{(g)}) \cap (\delta_{k,1r} \delta_{k,2c} = 1)), \quad (56)$$

dove abbiamo indicato con:  $\times$  l'operatore di prodotto matriciale;  $\mathbf{0}_v$  un vettore costituito da  $v$  valori identicamente pari a zero;  $\mathbf{A}_{RR}$  una matrice diagonale di dimensione ( $R \times R$ ) il cui  $i$ -esimo ( $i=1,...,R$ ) elemento sulla diagonale principale è dato da  $\tilde{N}_{(g),li} = \sum_{k \in s_{(g)}} d_k \delta_{k,li}$ ;  $\mathbf{A}_{RC}$  una matrice di

dimensione ( $R \times C$ ) il cui elemento che occupa la riga  $i$ -esima ( $i=1,...,R$ ) e la colonna  $j$ -esima ( $j=1,...,C$ ) è espresso da  $\tilde{N}_{(g),li,2j} = \sum_{k \in s_{(g)}} d_k \delta_{k,li} \delta_{k,2j}$  ;

$\mathbf{A}_{CC}$  una matrice diagonale di dimensione ( $C \times C$ ) il cui  $j$ -esimo ( $j=1,...,C$ ) elemento sulla diagonale principale è calcolato come  $\tilde{N}_{(g),2c} = \sum_{k \in s_{(g)}} d_k \delta_{k,2j}$

Dopo alcuni passaggi, indicando con

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

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

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

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

$$\mathbf{B}_{RR} = \mathbf{A}_{RR}^{-1} + \mathbf{A}_{RR}^{-1} \mathbf{A}_{RC} \left( \mathbf{A}_{CC} - \mathbf{A}_{RC}' \mathbf{A}_{RR}^{-1} \mathbf{A}_{RC} \right)^{-1} \mathbf{A}_{RC}' \mathbf{A}_{RR}^{-1} ;$$

$b_{(RC),ic}$  l'elemento nella riga  $i$ -esima e nella colonna  $c$ -esima della matrice

$$\mathbf{B}_{RC} = -\mathbf{A}_{RR}^{-1} \mathbf{A}_{RC} \left( \mathbf{A}_{CC} - \mathbf{A}_{RC}' \mathbf{A}_{RR}^{-1} \mathbf{A}_{RC} \right)^{-1} ;$$

$b_{(CC),jc}$  l'elemento nella riga  $j$ -esima e nella colonna  $c$ -esima della matrice

$$\mathbf{B}_{CC} = -\left( \mathbf{A}_{CC} - \mathbf{A}_{RC}' \mathbf{A}_{RR}^{-1} \mathbf{A}_{RC} \right)^{-1} \mathbf{A}_{RC}' \mathbf{A}_{RR}^{-1} ;$$

$b_{(RC),rj}$  l'elemento nella riga  $r$ -esima e nella colonna  $j$ -esima della matrice  $\mathbf{B}_{RC}$ .



## A.3 La costruzione dei data-set di input per definire i gruppi di riferimento

Nella presente appendice sono trattati con maggiore approfondimento i criteri di costruzione del *data-set* di input ed, in particolare, le alternative possibili per definire le variabili POP\_PIAN e le variabili TX<sub>j</sub> e X<sub>j</sub> (j=1, ..., J) in relazione al processo di stima che per calcolare i coefficienti finali di output. Tali criteri sono stati introdotti nel paragrafo 1.2. (Sez. II).

### A.3.1 Costruzione dei gruppi di riferimento: caso I

*Gruppi di riferimento definiti su sottopopolazioni pianificate ottenute marginalizzando su alcune variabili che contribuiscono a definire la stratificazione e con variabili ausiliarie  $x$  quantitative o qualitative dicotomiche*

Questo primo approfondimento sulla costruzione del data-set di input prevede due ipotesi di base:

- la prima richiede che la variabile di stratificazione sia multivariata<sup>24</sup> e che le sottopopolazioni pianificate, definite come aggregazioni di strati, siano il risultato di un processo di aggregazione rispetto ad una o più variabili che identificano gli strati stessi;
- la seconda ipotesi suppone che le variabili qualitative interessate dal processo di calibrazione siano dicotomiche del tipo presenza/assenza, sì/no, 0/1, ecc..

---

<sup>24</sup> Per variabile multivariata si intende che ciascuna modalità può essere definita come la combinazione delle modalità di due o più variabili

Nella prima ipotesi rientrano anche le strategie di campionamento in cui le sottopopolazioni pianificate coincidono con gli strati. In tal caso non è necessario distinguere le variabili di stratificazione tra semplici e multivariate.

Per rendere chiari gli aspetti sollevati si consideri l'esempio di seguito descritto.

#### *ESEMPIO A.3.1:*

*Sia dato un campione d'individui stratificato sulla base di una variabile che è ottenuta dalla combinazione di quattro variabili descritte nella tabella A.3.1.*

**Tabella A.3.1 - Variabili dell'esempio che descrivono la stratificazione**

<i>Variabili che definiscono la stratificazione</i>	<i>Simbolo variabile</i>	<i>Numero di modalità</i>	<i>Simbolo numero di modalità</i>	<i>Modalità</i>
Sesso	$s_1$	2	$S_1$	uomo; donna
Classe di età	$s_2$	4	$S_2$	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre
Stato civile	$s_3$	2	$S_3$	sposato; non sposato
Ripartizione geografica	$s_4$	3	$S_4$	nord; centro; sud

*Si considerino anche tre variabili di post-stratificazione (che non rientrano nella definizione degli strati del disegno), descritte nella tabella A.3.2, sulle quali sono noti i totali di alcune variabili ausiliarie utilizzate per definire lo stimatore di ponderazione vincolata adottato.*

**Tabella A.3.2 - Variabili dell'esempio che definiscono sottopopolazioni non pianificate (variabili di post-stratificazione)**

<i>Variabili di post-stratificazione</i>	<i>Simbolo</i>	<i>Numero di modalità</i>	<i>Simbolo numero di modalità</i>	<i>Modalità</i>
Settore di attività economica in cui lavora l'individuo	$v_1$	3	$Q_1$	agricoltura; industria; terziario;
Professione	$v_2$	4	$Q_2$	operaio; impiegato; dirigente; altro
Titolo di studio	$v_3$	4	$Q_3$	licenza elementare; licenza media; diploma di scuola superiore; laurea universitaria

*Si abbiano inoltre quattro variabili ausiliarie, definite nella tabella A.3.3, utilizzate*

*nello stimatore di ponderazione vincolata, per le quali si conoscono i totali di popolazione su alcune partizioni in gruppi di riferimento della popolazione.*

**Tabella A.3.3 - Variabili dell'esempio che presentano dei totali noti a livello di sottopopolazioni**

Variabili ausiliarie	Simbolo	Numero di modalità	Modalità
Indicatore di presenza dell'unità nella sottopopolazione.	$x_1$	2	appartenente alla sottopopolazione (1), non appartenente alla sottopopolazione (0). Per come viene definito il data-set la variabile è sempre pari a "1".
Indicatore di proprietà dell'abitazione	$x_2$	2	proprietario (1), non proprietario (0)
Numero di figli	$x_3$	-	-
Reddito individuale	$x_4$	-	-

*Infine, si considerino cinque differenti partizioni della popolazione in gruppi di riferimento, descritte nella tabella A.3.4, in cui sono noti i totali di popolazione per alcune delle variabili ausiliarie introdotte nella tabella A.3.3.*

**Tabella A.3.4 – Descrizione delle partizioni in gruppi di riferimento prese in considerazione nell'esempio**

Partizioni	Simbolo	Variabili che definiscono i gruppi di riferimento della partizione	Numero dei gruppi di riferimento nella partizione	Variabili ausiliarie per le quali si hanno i totali noti
Prima partizione	$P_1$	$s_1, s_2, s_3, s_4$	$D_1$	$x_1$
Seconda partizione	$P_2$	$s_1, s_2, v_2$	$D_2$	$x_2, x_4$
Terza partizione	$P_3$	$s_1, s_2, s_4, v_3$	$D_3$	$x_3$
Quarta partizione	$P_4$	$s_1, s_2, v_1$	$D_4$	$x_3, x_4$
Quinta partizione	$P_5$	$s_1, s_2, s_4, v_1$	$D_5$	$x_4$

*Per rendere chiaro quali sono le informazioni contenute nella tabella A.3.4 si osservi, ad esempio, la prima riga relativa alla prima partizione in gruppi di riferimento. Ciascun gruppo di riferimento di questa partizione è identificato da una particolare combinazione delle modalità di tutte le variabili che definiscono gli strati del disegno. Per i gruppi di questa prima partizione il totale utilizzato, a livello di stimatore di ponderazione vincolata, è il totale della popolazione.*

*La seconda partizione (seconda riga) è costituita dai gruppi di riferimento identificati dall'incrocio delle modalità della variabile sesso, classe di età e della professione.*



*In tali gruppi sono noti il totale di sottopopolazione degli individui possessori di una abitazione e il totale di sottopopolazione dei redditi individuali.*

*Per concludere questa breve descrizione delle caratteristiche delle cinque partizioni si può osservare che: la prima partizione presenta come sottopopolazioni pianificate i singoli strati; le restanti partizioni si basano, invece, su sottopopolazioni pianificate ricavate marginalizzando su una o più variabili che definiscono la stratificazione. In particolare, nella seconda partizione si marginalizza sulle variabili  $s_3, s_4$ , nella terza partizione si marginalizza sulla variabile  $s_3$ , e così via nelle altre due partizioni.*

*L'utente per indicare al software quali sono le partizioni in gruppi di riferimento della popolazione obiettivo, deve agire sulla definizione delle modalità della variabile POP\_PLAN, sulla costruzione di un certo numero di variabili  $TX_j$  e  $X_j$  (si veda il paragrafo 1.2, Sez. II) e sulla definizione dei valori che possono assumere queste ultime. A tale scopo si possono adottare una delle tre alternative introdotte nel paragrafo 1.2., Sez. II. Nelle tabelle A.3.5, A.3.6 e A.3.7 è descritto come costruire il data set "dati campionari".*

**Tabella A.3.5 - Costruzione del data-set "dati campionari" secondo lo schema A**

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$ (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili $X_j$
Numero Modalità	POP_PIAN	1	1	Nessuna variabile identifica POP_PIAN
		$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_1$ sulla partizione $P_1$	48 $X_1, \dots, X_{48}$	$\times$ $S_1 \times S_2 \times S_3 \times S_4$	$S_1, S_2, S_3, S_4$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_2$ sulla partizione $P_2$	32 $X_{49}, \dots, X_{80}$	$S_1 \times S_2 \times Q_2$	$S_1, S_2, V_2$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_3$	96 $X_{81}, \dots, X_{176}$	$S_1 \times S_2 \times S_4 \times Q_3$	$S_1, S_2, S_4, V_3$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_4$	24 $X_{177}, \dots, X_{200}$	$S_1 \times S_2 \times Q_1$	$S_1, S_2, V_1$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_2$	32 $X_{201}, \dots, X_{232}$	$S_1 \times S_2 \times Q_2$	$S_1, S_2, V_2$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_4$	24 $X_{233}, \dots, X_{256}$	$S_1 \times S_2 \times Q_1$	$S_1, S_2, V_1$
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_5$	72 $X_{257}, \dots, X_{328}$	$S_1 \times S_2 \times S_4 \times Q_1$	$S_1, S_2, S_4, V_1$
Numero totale di variabili $X_j$ nel data-set di input		328 $X_1, \dots, X_{328}$		

**Tabella A.3.6 - Costruzione del data-set "dati campionari" secondo lo schema B**

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$ (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili $X_j$
Numero Modalità	POP_PIAN	8	$S_1 \times S_2$	$s_1, s_2$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_1$ sulla partizione $P_1$	6 $X_1, \dots, X_6$	$S_3 \times S_4$	$s_3, s_4$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_2$ sulla partizione $P_2$	4 $X_7, \dots, X_{10}$	$Q_2$	$v_2$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_3$	12 $X_{11}, \dots, X_{22}$	$S_4 \times Q_3$	$s_4, v_3$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_4$	3 $X_{23}, \dots, X_{25}$	$Q_1$	$v_1$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_2$	4 $X_{26}, \dots, X_{28}$	$Q_2$	$v_2$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_4$	3 $X_{30}, \dots, X_{32}$	$Q_1$	$v_1$
	$X_j$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Numero variabili	Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_5$	9 $X_{33}, \dots, X_{41}$	$S_4 \times Q_1$	$s_4, v_1$
Numero totale di variabili $X_j$ nel data-set di input		41 $X_1, \dots, X_{41}$		

**Tabella A.3.7 - Costruzione del data-set "dati campionari" secondo lo schema C**

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$	Numero delle modalità della variabile POP_PIAN e numero delle variabili $X_j$ (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili $X_j$
Numero Modalità	POP_PIAN	2 (oppure 4)	$S_1$  (oppure $S_2$ )	$s_1$  (oppure $s_2$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_1$ sulla partizione $P_1$	$\Downarrow$ 24 $X_1, \dots, X_{24}$ (oppure 12 $X_1, \dots, X_{12}$ )	$\Downarrow$ $S_2 \times S_3 \times S_4$  (oppure $S_1 \times S_3 \times S_4$ )	$\Downarrow$ $s_2, s_3, s_4$  (oppure $s_1, s_3, s_4$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_2$ sulla partizione $P_2$	16 $X_{25}, \dots, X_{40}$ (oppure 8 $X_{13}, \dots, X_{20}$ )	$S_2 \times Q_2$  (oppure $S_1 \times Q_2$ )	$s_2, v_2$  (oppure $s_1, v_2$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_3$	48 $X_{41}, \dots, X_{88}$ (oppure 24 $X_{21}, \dots, X_{44}$ )	$S_2 \times S_4 \times Q_3$  (oppure $S_1 \times S_4 \times Q_3$ )	$s_2, s_4, v_3$  (oppure $s_1, s_4, v_3$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_3$ sulla partizione $P_4$	12 $X_{89}, \dots, X_{100}$ (oppure 6 $X_{45}, \dots, X_{50}$ )	$S_2 \times Q_1$  (oppure $S_1 \times Q_1$ )	$s_2, v_1$  (oppure $s_1, v_1$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_2$	16 $X_{101}, \dots, X_{116}$ (oppure 8 $X_{51}, \dots, X_{58}$ )	$S_2 \times Q_2$  (oppure $S_1 \times Q_2$ )	$s_2, v_2$  (oppure $s_1, v_2$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_4$	12 $X_{117}, \dots, X_{128}$ (oppure 6 $X_{59}, \dots, X_{64}$ )	$S_2 \times Q_1$  (oppure $S_1 \times Q_1$ )	$s_2, v_1$  (oppure $s_1, v_1$ )
Numero variabili	$X_j$ Per tenere conto dei totali della variabile $x_4$ sulla partizione $P_5$	36 $X_{129}, \dots, X_{164}$ (oppure 18 $X_{65}, \dots, X_{82}$ )	$S_2 \times S_4 \times Q_1$  (oppure $S_1 \times S_4 \times Q_1$ )	$s_2, s_4, v_1$  (oppure $s_1, s_4, v_1$ )
Numero totale di variabili $X_j$ nel data-set di input		164 $X_1, \dots, X_{164}$ (oppure 82 $X_1, \dots, X_{82}$ )		

*Le informazioni contenute nelle tre tabelle belle sono le seguenti:*

Schema A (tabella A.3.5);

- 1° la variabile POP\_PLAN ha una sola modalità. Tutti i record presentano un valore costante della variabile;*
- 2° sono presenti le variabili X1, ..., X328. L'insieme di queste variabili è suddiviso in sette sottoinsiemi:*
- 3° sottoinsieme che raggruppa le variabili X1, ..., X48: queste variabili identificano i valori della variabile  $x_1$  sulla partizione  $P_1$ ; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1, s_2, s_3, s_4$  che si presenta sul record corrispondente;*
- 4° sottoinsieme che raggruppa le variabili X49, ..., X80: queste variabili identificano i valori della variabile  $x_2$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili  $s_1, s_2, v_2$  che si presenta sul record corrispondente;*
- 5° sottoinsieme che raggruppa le variabili X81, ..., X176: queste variabili identificano i valori della variabile  $x_3$  sulla partizione  $P_3$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1, s_2, s_4, v_3$  che si presenta sul record corrispondente;*
- 6° sottoinsieme che raggruppa le variabili X177, ..., X200: queste variabili identificano i valori della variabile  $x_3$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1, s_2, v_1$  che si presenta sul record corrispondente;*
- 7° sottoinsieme che raggruppa le variabili X201, ..., X232: queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non*

*nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1$ ,  $s_2$ ,  $v_2$  che si presenta sul record corrispondente;*

- 8° *sottoinsieme che raggruppa le variabili X233, ..., X256: queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1$ ,  $s_2$ ,  $v_1$ , che si presenta sul record corrispondente;*
- 9° *sottoinsieme che raggruppa le variabili X257, ..., X328: queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_5$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_1$ ,  $s_2$ ,  $s_4$ ,  $v_1$  che si presenta sul record corrispondente;*

Schema B (tabella A.3.6);

- 1° *le modalità assunte dalla variabile POP\_PLAN identificano le differenti combinazioni delle modalità delle variabili  $s_1$ ,  $s_2$ . In particolare, ciascun record presenta sulla variabile POP\_PLAN la modalità che identifica la combinazione di  $s_1$ ,  $s_2$  presente nel record stesso.*
- 2° *sono presenti le variabili X1, ..., X41. L'insieme di queste variabili è suddiviso in sette sottoinsiemi:*
- 3° *sottoinsieme che raggruppa le variabili X1, ..., X6: queste variabili identificano i valori della variabile  $x_1$  sulla partizione  $P_1$ ; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_3$ ,  $s_4$  che si presenta sul record corrispondente;*
- 4° *sottoinsieme che raggruppa le variabili X7, ..., X10: queste variabili identificano i valori della variabile  $x_2$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili  $v_2$  che si presenta sul record corrispondente;*
- 5° *sottoinsieme che raggruppa le variabili X11, ..., X22: queste variabili identi-*

ficano i valori della variabile  $x_3$  sulla partizione  $P_3$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_4, v_3$  che si presenta sul record corrispondente;

- 6° sottoinsieme che raggruppa le variabili  $X23, \dots, X25$ : queste variabili identificano i valori della variabile  $x_3$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $v_1$  che si presenta sul record corrispondente;
- 7° sottoinsieme che raggruppa le variabili  $X26, \dots, X29$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $v_2$  che si presenta sul record corrispondente;
- 8° sottoinsieme che raggruppa le variabili  $X30, \dots, X32$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $v_1$  che si presenta sul record corrispondente;
- 9° sottoinsieme che raggruppa le variabili  $X33, \dots, X41$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_5$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_4, v_1$  che si presenta sul record corrispondente.

Schema C (tabella A.3.7);

Relativamente allo schema C la tabella rileva l'esistenza di due possibili alternative. La prima definisce le modalità della variabile POP\_PLAN in base alla variabile  $s_1$ , la seconda, invece, sulla variabile  $s_2$ . Descrivendo la prima delle due alternative si ha che:

- 1° le modalità assunte dalla variabile POP\_PLAN identificano (possono anche coincidere) le modalità delle variabili  $s_1$ . In particolare, ciascun record presenta sulla variabile POP\_PLAN la modalità che identifica la modalità di  $s_1$  che si presenta nel record stesso.
- 2° sono presenti le variabili  $X1, \dots, X164$ . L'insieme di queste variabili è suddiviso in sette sottoinsiemi:
- 3° sottoinsieme che raggruppa le variabili  $X1, \dots, X24$ : queste variabili identificano i valori della variabile  $x_1$  sulla partizione  $P_1$ ; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, s_3, s_4$  che si presenta sul record corrispondente;
- 4° sottoinsieme che raggruppa le variabili  $X25, \dots, X40$ : queste variabili identificano i valori della variabile  $x_2$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili  $s_2, v_2$  che si presenta sul record corrispondente;
- 5° sottoinsieme che raggruppa le variabili  $X41, \dots, X88$ : queste variabili identificano i valori della variabile  $x_3$  sulla partizione  $P_3$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, s_4, v_3$  che si presenta sul record corrispondente;
- 6° sottoinsieme che raggruppa le variabili  $X89, \dots, X100$ : queste variabili identificano i valori della variabile  $x_3$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, v_1$ , che si presenta sul record corrispondente;
- 7° sottoinsieme che raggruppa le variabili  $X101, \dots, X116$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_2$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, v_2$  che si presenta sul record corrispondente;



8° sottoinsieme che raggruppa le variabili  $X_{117}, \dots, X_{128}$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_4$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, v_1$  che si presenta sul record corrispondente;

9° sottoinsieme che raggruppa le variabili  $X_{129}, \dots, X_{164}$ : queste variabili identificano i valori della variabile  $x_4$  sulla partizione  $P_5$ ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili  $s_2, s_4, v_1$  che si presenta sul record corrispondente.

Per rendere più generale la descrizione vista nell'esempio A.3.1 dei tre schemi di costruzione di un *data-set* di input, è necessario definire una simbologia, in parte già introdotta nell'esempio stesso, per identificare le variabili che rappresentano gli strati (tabella A.3.8), i post-strati (tabella A.3.9) e le variabili di cui si usano i totali noti a livello di stimatore (tabella A.3.10). Relativamente a queste ultime, si considerano, per il momento, le variabili quantitative e le variabili qualitative dicotomiche del tipo presenza/assenza, sì/no, 0/1.

**Tabella A.3.8 – Definizione simbolica delle variabili che identificano uno strato**

Variabile	$S_1$	...	$S_a$	...	$S_A$
Numero di modalità	$S_1$	...	$S_a$	...	$S_A$

**Tabella A.3.9 – Definizione simbolica delle variabili di post-stratificazione**

Variabile	$V_1$	...	$V_b$	...	$V_B$
Numero di modalità	$Q_1$	...	$Q_b$	...	$Q_B$

**Tabella A.3.10 – Definizione simbolica delle variabili ausiliarie di cui si utilizzano i totali noti a livello di stimatore**

Variabile	$x_1$	...	$x_t$	...	$x_T$
-----------	-------	-----	-------	-----	-------

In base alla notazione presentata nelle tabelle A.3.8 e A.3.9, una generica partizione  $P_i$  ( $i=1, \dots, I$ ), è definibile da un sottoinsieme  $\underline{s}^i$  composto da alcune delle variabili  $s_a$  ( $a=1, \dots, A$ ) e da una variabile di post-stratificazione  $v^i$  che coincide con una delle variabili  $v_b$  ( $b=1, \dots, B$ ). Tale partizione, come è illustrato nella tabella A.3.11, è composta da  $\underline{S}^i \times Q^i$  gruppi di riferimento, dove  $\underline{S}^i$  è il numero di combinazioni di modalità delle variabili contenute nel sottoinsieme  $\underline{s}^i$ , mentre  $Q^i$  è il numero di modalità di  $v^i$ .

**Tabella A.3.11 – Descrizione simbolica delle partizioni in gruppi di riferimento di una popolazione oggetto d'indagine**

Indicatore di partizione	$P_1$	...	$P_i$	...	$P_I$
Insieme di variabili di stratificazione che identificano la partizione	$\underline{s}^1$	...	$\underline{s}^i$	...	$\underline{s}^I$
Numero delle combinazioni di modalità delle variabili di stratificazione che identificano la partizione	$\underline{S}^1$	...	$\underline{S}^i$	...	$\underline{S}^I$
Variabile di post-stratificazione che identifica la partizione	$v^1$	...	$v^i$	...	$v^I$
Numero delle modalità della variabile di post-stratificazione che identifica la partizione	$Q^1$	...	$Q^i$	...	$Q^I$
Numero dei gruppi di riferimento della partizione	$\underline{S}^1 \times Q^1$	...	$\underline{S}^i \times Q^i$	...	$\underline{S}^I \times Q^I$

Dati questi elementi, si indichi con  $\underline{s}$  il sottoinsieme delle variabili di stratificazione che sono contenute in tutti gli insiemi  $\underline{s}^i$ . Inoltre sia  $\underline{S}$  il numero di combinazioni delle modalità delle variabili in  $\underline{s}$ . Pertanto, per la generica partizione  $P_i$ , il numero dei gruppi di riferimento si può denotare con il prodotto  $\underline{S} \times \underline{\bar{S}}^i \times Q^i$  in cui  $\underline{\bar{S}}^i$  è il numero delle combinazioni delle modalità dell'insieme di variabili  $\underline{\bar{s}}^i$  incluse in  $\underline{s}^i$  ed escluse da  $\underline{s}$ , avendo, quindi,  $\underline{s}^i = \underline{s} \cup \underline{\bar{s}}^i$ .

Considerata la simbologia sopra introdotta, è possibile, allora, dare una struttura generale per definire lo schema A e lo schema B (si veda tabella A.3.12).

Per impostare il *data-set* di input secondo lo schema C è necessario definire con  ${}^c\underline{s}$  e  ${}^c\underline{\bar{s}}$  due sottoinsiemi di variabili tra loro disgiunti la cui unione riporta ad  $\underline{s}$ . Si indichi con  ${}^c\underline{S}$  il numero delle combinazioni delle

modalità delle variabili in  ${}^c\underline{s}$  e con  $\bar{c}\underline{S}$  il numero delle combinazioni delle modalità delle variabili in  $\bar{c}\underline{s}$ . Dunque, attraverso questa nuova notazione il numero dei gruppi di riferimento per la generica partizione  $P_i$  è data dal prodotto  ${}^c\underline{S} \times \bar{c}\underline{S} \times \bar{S}^i \times Q^i$ . Come è illustrato nella tabella A.3.12 la scissione di  $\underline{s}$  nei due sottoinsiemi, consente l'attuazione dello schema C.

**Tabella A.3.12 – Descrizione degli schemi di costruzione del data-set di input: definizione del numero di modalità della variabile POP\_PIAN e del numero di variabili  $X_j$**

SCHEMA		Numero delle modalità della variabile POP_PIAN		Numero di variabili $X_j$ per ogni variabile $X_t$ definita in $P_1$	Numero di variabili $X_j$ per ogni variabile $X_t$ definita in $P_i$	Numero di variabili $X_j$ per ogni variabile $X_t$ definita in $P_i$
A		1	$\Rightarrow$	$\underline{S} \times \bar{S}^1 \times Q^1$	$\underline{S} \times \bar{S}^i \times Q^i$	$\bar{S}^i \times Q^i$
B		$\underline{S}$	$\Rightarrow$	$\bar{S}^1 \times Q^1$	$\bar{S}^i \times Q^i$	$\bar{S}^i \times Q^i$
C	Due alternative	${}^c\underline{S}$	$\Rightarrow$	$\bar{c}\underline{S} \times \bar{S}^i \times Q^i$	$\bar{c}\underline{S} \times \bar{S}^i \times Q^i$	$\bar{c}\underline{S} \times \bar{S}^i \times Q^i$
		$\bar{c}\underline{S}$		${}^c\underline{S} \times \bar{S}^i \times Q^i$	${}^c\underline{S} \times \bar{S}^i \times Q^i$	${}^c\underline{S} \times \bar{S}^i \times Q^i$

Dalla tabella si evidenziano alcune considerazioni già espresse a conclusione dell'esempio A1: in primo luogo lo schema B è inapplicabile quando  $\underline{s}$  è un insieme vuoto (o, in altri termini lo schema B coincide con lo schema A); in secondo luogo lo schema C è inapplicabile quando  $\underline{s}$  contiene una sola variabile (o, in altri termini lo schema C coincide con lo schema B).

### A.3.2 Costruzione dei gruppi di riferimento: caso II

*Gruppi di riferimento nel caso di sottopopolazioni pianificate ottenute non marginalizzando la variabile di stratificazione multivariata e con variabili qualitative ausiliarie  $x$  di tipo non dicotomico*

Gli schemi illustrati nella tabella A.3.12 non comprendono tutti i tipi di partizioni in gruppi di riferimento e tutti i tipi di variabili ausiliarie che possono essere utilizzate per definire uno stimatore di calibrazione. Infatti, nel descrivere l'impostazione del *data-set* di input si è fatto riferimento a due ipotesi restrittive che non sempre si verificano nella pianificazione di una strategia di campionamento: la prima ipotesi prevede che il processo di aggregazione degli strati per definire le sottopopolazioni pianificate, avvenga marginalizzando rispetto ad una o più variabili che individuano gli stessi strati; la seconda suppone che le variabili qualitative  $x$  siano dicotomiche, del tipo presenza/assenza, sì/no, 0/1, ecc..

Di seguito sono illustrati i passi necessari per impostare il *data-set* di input quando le ipotesi precedenti non sono proprie della strategia campionaria adottata dall'utente.

Per comprendere quali sono le implicazioni che intervengono quando non si verifica la prima ipotesi è utile considerare il seguente esempio:

#### ESEMPIO A.3.2

- *Sia dato un disegno campionario in cui la stratificazione avviene su una variabile multivariata ottenuta dalle variabili sesso (2 modalità: uomini - U; donne - D) e classe di età (4 modalità: 0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre). Su tale stratificazione si può effettuare un primo tipo di aggregazione degli strati marginalizzando sulla classe di età e formando, pertanto, due gruppi di strati: il primo identificato dalla modalità U, il secondo dalla modalità D.*
- *Con questa stratificazione la strategia campionaria potrebbe, tuttavia, presentare un secondo tipo di aggregazione degli strati che coinvolge l'unione di alcune modalità all'interno di una variabile che identifica gli strati senza procedere alla marginalizzazione rispetto ad una specifica variabile. Ciò avviene, ad esempio, aggregando gli strati identificati dalle modalità 0-14 anni e 15-34 anni della variabile classe di età, ottenendo, dunque, sei gruppi di strati, identificati esattamente da: U e 0-34 anni, D e 0-34 anni, U e 35-54 anni, D e 35-54 anni, U e 55 anni e oltre, D e 55 anni e oltre.*

*L'aggregazione degli strati che non prevede una marginalizzazione rispetto a variabili che identificano gli strati stessi non presenta particolari problemi dal punto di vista operativo. Bisogna, tuttavia, distinguere due casi:*

- *il primo prevede che la procedura di aggregazione degli strati è la stessa su tutte le partizioni considerate;*
- *il secondo permette di avere differenti procedure di aggregazione che cambiano al cambiare delle partizioni. Riprendendo l'esempio, la strategia campionaria si potrebbe presentare una prima partizione ottenuta aggregando degli strati con classe di età 0-14 anni e 15-34 anni e una seconda partizione in cui si aggregano fra loro gli strati con modalità 0-14 anni e 15-34 anni e gli strati con modalità 35-54 anni e 55 anni e oltre. In tutti i casi le aggregazioni avvengono per strati che presentano la stessa modalità della variabile sesso.*

*Facendo riferimento alla suddivisione delle variabili, in variabili che definiscono gli strati e variabili di post-stratificazione, necessaria per impostare i due archivi di input, si deve procedere in due modi differenti per i due casi:*

- *nel primo caso si sostituisce la variabile in cui avvengono le aggregazioni, con una nuova variabile le cui modalità sono aggregazioni delle modalità di quella originale. Così, se l'aggregazione degli strati è quella presentata nell'esempio e se questo criterio di aggregazione si ripete in tutte le partizioni previste dalla strategia campionaria, la variabile classe di età con quattro modalità (0-14 anni;15-34 anni;35-54 anni;55 anni e oltre) è sostituita nella definizione del data-set di input con una nuova variabile che presenta tre modalità (0-34 anni;35-54 anni;55 anni e oltre);*
- *nel secondo caso l'originale variabile di stratificazione non viene considerata nella formazione del data-set, mentre sono prese in considerazione tante nuove variabili di post-stratificazione per quante sono le differenti aggregazioni in strati. Ad esempio, considerando il punto ii nella costruzione dell'archivio di input si deve escludere la variabile classe di età come variabile che definisce gli strati e si devono inserire una prima nuova variabile di post-stratificazione con tre modalità (0-34;35-54;55 e oltre) e una seconda nuova variabile di post-stratificazione con due modalità (0-34;35 e oltre).*

Considerando ora il caso in cui le variabili qualitative inserite nel processo di calibrazione non sono dicotomiche (presenza/assenza; 0/1 ecc.) è necessario operare una loro preventiva trasformazione nella forma detta disgiuntiva completa.

Sia data per esempio la variabile “titolo di studio” con quattro modalità: “licenza elementare”; “licenza media”; “diploma di scuola superiore”; “laurea universitaria”. In questo caso la forma disgiuntiva completa della variabile definisce le seguenti quattro variabili dicotomiche: “il titolo di studio è la licenza elementare con modalità sì/no”; “il titolo di studio è la licenza media con modalità sì/no”; “il titolo di studio è diploma di scuola superiore con modalità sì/no”; “il titolo di studio è la laurea universitaria con modalità sì/no”.

Sulla base di queste quattro variabili saranno definite successivamente le variabili  $X_j$  secondo l’opportuno schema di impostazione dei due *data-set* di input.



## **BIBLIOGRAFIA**

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

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

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

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

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

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

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



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

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

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

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

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

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

Pagliuca, D. (a cura di), 2004b, Genesees v.3.0., Funzione Stime ed Errori Manuale utente ed aspetti metodologici, Tecniche e Strumenti, ISTAT, n. 3. (disponibile anche su sito: via internet (per utenti esterni all'istat): <http://www.istat.it/Metodologi/index.htm> (selezionare "Metodi e Software per indagini statistiche"); via intranet (per utenti istat): <http://intranet/> (selezionare: "Prodotti e Applicazioni on-line. Software Generalizzati" e da qui selezionare "MPS-E: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)").

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

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

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

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

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

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

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

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