SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXVI Ciclo – anno 2013/2014

Dipartimento di Statistica

Coerenza ed ottimalità delle stime calibrate su informazioni da indagini campionarie

Alessio Guandalini

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Il ricorso ad informazioni ausiliarie è particolarmente utile nelle indagini campionarie in quanto il loro impiego aiuta a migliorarne notevolmente la qualità.

Per informazioni ausiliarie si intendono tutte quelle informazioni che non rappresentano l'obiettivo primario dell'indagine. Queste possono essere già note nel frame da cui le unità campionarie vengono estratte o possono essere rilevate sulle unità campionarie e la loro media o il loro totale noto nella popolazione.

L'impiego di informazioni ausiliarie generalmente non comporta costi di rilevazione aggiuntivi e può avvenire a diversi livelli del processo di produzione di dati campionari. Ad esempio, in fase di progettazione dell'indagine, in fase di editing ed imputazione e in fase di stima.

Nel caso in cui le informazioni ausiliarie sono note nel frame da cui le unità sono estratte, queste possono essere adottate per definire disegni campionari più efficienti. Ad esempio, possono essere utilizzate per programmare un disegno stratificato in modo da costruire strati omogenei che presentino una variabilità minore rispetto alla variabile di interesse.

Quando, invece, si conoscono medie e totali delle variabili ausiliarie nella popolazione è possibile costruire campioni bilanciati (cfr. e.g. Gini e Galvani (1929); Deville e Tillé (2004)), ovvero, campioni in grado di riprodurre in maniera esatta, o approssimata, il totale (o la media) delle variabili ausiliarie considerate. Così viene soddisfatta una delle possibili definizioni di *rappresentatività* del campione, in quanto questo può essere visto come una miniatura della popolazione, ovviamente in riferimento alle sole variabili ausiliarie considerate. Le variabili ausiliarie, ovviamente, dovranno essere scelte in maniera opportuna tenendo conto del legame con la variabile di interesse.

L'impiego delle variabili ausiliarie risulta, inoltre, particolarmente utile nella fase di editing e di imputazione, in quanto possono essere utilizzate per correggere la distorsione generata dalla mancata risposta totale (Särndal e Lunström, 2005). Nella maggior parte dei casi reali, l'ipotesi di *ignorabilità* del meccanismo che genera la

mancata risposta, ovvero l'ipotesi di uguaglianza di comportamento tra chi partecipa all'indagine e chi non partecipa, non può essere valida. Il ricorso a variabili ausiliarie correlate con la mancata risposta in modelli o per definire le cosiddette *celle di ponderazione*, dunque, può aiutare a determinare dei fattori correttivi che consentono di recuperare la numerosità campionaria teorica.

Infine, le variabili ausiliarie possono essere inserite nel processo di stima attraverso il ricorso ai loro totali (o alle loro medie) note nella popolazione con l'obiettivo primario di migliorare l'efficienza.

Il lavoro svolto, si colloca all'interno di questo ultimo contesto, ovvero, l'impiego di variabili ausiliarie nel processo di stima.

In letteratura sono noti diversi stimatori che, sfruttando la conoscenza del totale delle variabili ausiliarie, migliorano l'efficienza delle stime prodotte. In particolar modo lo stimatore di regressione generalizzata ($GREG$ cfr., e.g., Cassel *e altri* (1979); Fuller (2002)) e, più recentemente, lo stimatore calibrato (CAL cfr., e.g. Deville e Särndal (1992); Deville *e altri* (1993); Särndal (2007)), utilizzano i totali (o le medie) delle variabili ausiliare note da una fonte amministrativa o censuaria per inglobare queste informazioni nel processo di stima e migliore l'efficienza delle stime prodotte. Una loro proprietà, particolarmente importante, per il prosieguo della trattazione, è quella di condurre a stime coerenti con i totali di controllo delle variabili ausiliarie.

Per entrambi questi stimatori l'inferenza sui parametri di interesse può essere svolta seguendo un approccio *design-based*, basato sul piano di campionamento, o un approccio *model-based*, basato su un modello di superpopolazione. In aggiunta può essere seguito un approccio intermedio tra questi due, *model-assisted*, in cui il modello di superpopolazione è assunto implicitamente per spiegare la relazione tra le variabili ausiliarie e la variabile di interesse ed è utilizzato per ottenere stime più efficienti, rispetto al disegno, di quelle dello stimatore di Horvitz-Thompson.

Nella loro definizione originaria, questi stimatori impiegano totali delle variabili ausiliarie, detti anche totali di controllo o totali noti, conosciuti da fonte amministrativa o censuaria, ovvero non affetta da errore campionario. Tuttavia, negli ultimi anni, è sempre maggiore la richiesta di inserire tra il gruppo di variabili ausiliarie considerate anche quelle di cui i totali non sono conosciuti da fonti al riparo da errori campionari, ma sono a loro volta delle stime campionarie. In questo caso, dunque, i totali di controllo saranno affetti da errore campionario e per questo detti totali di controllo campionari.

I motivi per cui si presenta la necessità di utilizzare anche variabili ausiliarie i

cui totali sono stime ottenute da indagini campionarie sono diversi e validi.

Il ricorso a totali di controllo campionari rende possibile inserire delle informazioni importanti per migliorare l'efficienza della stima che, altrimenti, non potrebbero essere considerate. Un esempio pratico, che è stato studiato nelle applicazioni, è dato dall'impiego di informazioni ausiliarie su condizione occupazionale e titolo di studio per migliorare le stime sui redditi. La forte relazione che lega la variabile reddito con quella del titolo di studio e della condizione professionale è nota, ma, non essendoci fonti amministrative dalle quali ricavare i totali aggiornati di queste variabili, per inserirle nel processo di stima, è necessario ricorrere a totali di controllo campionari.

Particolarmente diffuso, sopratutto nel caso in cui si applica un disegno campionario ruotato, è l'impiego di totali stimati in occasioni precedenti della stessa indagine. E' dimostrato, infatti, che l'uso di queste informazioni ausiliarie negli stimatori, che vengono definiti in questo caso *composite estimator*, porta a delle stime di gran lunga più efficienti e meno volatili (Wolter, 1979; Singh *e altri*, 2001; Fuller e Rao, 2001).

Un motivo ulteriore, particolarmente sentito all'interno degli Istituti nazionali di statistica, è l'esigenza di costruire un sistema di indagini coerenti. Contemporaneamente, infatti, vengono svolte dallo stesso Istituto diverse indagini relative alla stessa popolazione e allo stesso periodo di riferimento che producono stime per gli stessi parametri. La pubblicazione di stime diverse per lo stesso parametro relativo allo stesso aggregato e allo stesso periodo rappresenta un grande problema per l'Istituto e soprattutto per i fruitori dell'informazione statistica. Il ricorso a totali di controllo campionari può fornire in questo contesto una soluzione semplice e convincente per ovviare a questo problema.

Dal 2015 in Italia avverrà il passaggio dal censimento nell'accezione classica del termine, ovvero contraddistinto dalla caratteristica di universalità, al censimento "campionario". Il numero di totali di controllo senza errore campionario così si ridurrà. Ad esempio, tutti i totali di controllo relativi ai sistemi locali del lavoro, che in alcuni casi vengono utilizzati, diventeranno totali di controllo campionari. Inoltre soprattutto a distanza di un paio di anni dallo svolgimento della rilevazione censuaria, i totali desunti dal censimento possono essere obsoleti e quindi meno affidabili di stime campionarie ottenute nello stesso periodo in cui l'uso di questi totali è richiesto.

I totali di controllo campionari vengono già ampiamente utilizzati per migliorare le stime di piccole aree (Rao, 2003). Generalmente, infatti, nelle stime per piccole aree vengono impiegati degli stimatori indiretti, nel senso che utilizzano informazioni

all'esterno del dominio, ad esempio stime per *grandi* aree, per migliorare le stime all'interno del dominio, ovvero della piccola area.

Infine, un altro motivo che spinge al ricorso a totali noti campionari sono gli errori non campionari che possono affliggere le fonti amministrative (cfr. Lessler e Kalsbeek, 1992; Nicolini *e altri*, 2013). Queste, infatti, possono essere affette da errori di sotto-copertura, duplicazione, o in generale non essere aggiornate e, quindi, in alcuni casi, portare in dote un errore la cui dimensione non si conosce e che generalmente viene ignorato. Di contro, utilizzando i totali di controllo campionari si è consapevoli di commettere un errore campionario di cui si conosce l'entità, e come vedremo, di cui è possibile valutare le conseguenze nel processo di stima.

Da questa serie di motivi, appare chiaro come l'impiego di totali noti campionari sia un problema di particolare importanza ed attualità per quel che riguarda il processo di stima delle rilevazioni campionarie.

Preso atto della necessità di confrontarsi con stimatori che ricorrono a variabili ausiliari i cui totali di controllo sono campionari, gli interrogativi che sorgono sono principalmente due. Il primo è relativo all'impatto che l'errore campionario di questi totali può avere sulle stime, e quindi se il ricorso a questo tipo di informazioni ausiliarie sia conveniente oppure se, quando i totali non sono noti da fonti amministrative o censuarie, è preferibile farne a meno. Il secondo, invece, è relativo al modo "ottimale" in cui queste informazioni possono essere utilizzate per ottenere il massimo beneficio possibile in termini di efficienza delle stime.

Abbiamo, dunque, individuato due possibili punti di vista da cui è possibile affrontare il problema che corrispondono a due diversi obiettivi perseguibili in questo contesto.

La ricerca della coerenza, rappresenta un punto di vista più vicino a quello degli Istituti nazionali di statistica in quanto si ricollega alla necessità di costruire un sistema di indagini integrato. L'obiettivo principale è rappresentato dal rispetto della condizione di coerenza con totali di controllo campionari, così come avviene di norma per quelli non campionari. Il problema metodologico principale in questo contesto è rappresentato dalla determinazione dell'errore campionario ed, in particolar modo, dalla determinazione della quantità di errore aggiuntiva dovuta all'impiego di totali noti campionari (Dever e Valliant, 2010; Berger *e altri*, 2009).

La ricerca dell'ottimalità, che rappresenta l'altro punto di vista, è relativa all'impiego delle informazioni ausiliarie per sfruttarne al meglio la loro capacità informativa condizionatamente all'errore campionario dei loro totali. In questo ambito, dunque, il problema metodologico è rappresentato dalla minimizzazione dell'erro-

re campionario delle stime che viene raggiunta allentando in maniera opportuna la condizione di coerenza.

I due punti di vista, quindi, sono inconcigliabili e perseguire la coerenza implica, in genere, la rinuncia alla condizione di ottimalità e viceversa.

La struttura del lavoro segue, dunque, questa schema. Nel primo capitolo, verranno introdotti i concetti chiave per il proseguo della trattazione e, in particolar modo, presentati lo stimatore di regressione generalizzata e lo stimatore calibrato nella loro definizione classica, ovvero con l'impiego di totali di controllo non-campionari.

Nel secondo capitolo verrà affrontato il problema dell'impiego di totali noti campionari focalizzando l'attenzione sulla proprietà di coerenza. In particolare, dopo una breve illustrazione della metodologia *Repeated weighting* messa a punto dall'Istituto nazionale di statistica olandese per costruire un sistema integrato di indagini, viene proposto uno stimatore che soddisfa la condizione di coerenza per i totali di controllo campionari. Soprattutto, sono proposti i relativi stimatori della varianza campionaria sia nel caso in cui le stime dei totali sono ricavate da indagini indipendenti rispetto a quella per cui si stanno producendo le stime, sia in caso di totali stimati da indagini dipendenti.

Nel terzo capito viene affrontato il problema dal punto di vista dell'ottimalità. Viene proposto uno stimatore che, allentando o restringendo in base ad opportuni parametri la condizione di coerenza rispetto ai totali noti campionari, è in grado di avere varianza minima rispetto agli altri stimatori che utilizzano le stesse variabili. L'espressione di questo stimatore è derivata per alcuni casi di particolare interesse pratico.

Nel capitolo quarto, vengono illustrati in maniera approfondita i comportamenti degli stimatori proposti rispetto ad altri stimatori già noti in letteratura. Attraverso uno studio di simulazione verranno, infatti, studiate le proprietà degli stimatori tenendo sotto controllo il legame tra la variabile di interesse e le variabili ausiliarie. Mentre, attraverso l'applicazione a dati reali verrà illustrato e spiegato come l'uso di totali noti campionari può avere impatti diversi sulle stime relative a diversi domini. Questa parte assume notevole importanza, in quanto, si mette in evidenza come l'intervento del ricercatore e la conoscenza dell'indagine oltre che dei dati sono fondamentali nella decisione dell'impiego di questo tipo di informazioni ausiliarie.

Le applicazioni riguarderanno il caso di indagini indipendenti e dipendenti. Per il caso di indagini indipendenti sarà studiato il caso della componente italiana della rilevazione sui redditi e le condizioni di vita (It-Silc) che utilizza in maniera sistematica totali su condizione lavorativa e livello di istruzione stimati dalla rilevazione

sulle forze di lavoro (RCFL). Per il caso di indagini dipendenti, invece, sarà studiato il caso delle stime mensili della rilevazione sulle forze di lavoro che utilizzano totali stimati sui campioni di tre mesi e dodici mesi prima della stessa indagine. Infine, nelle conclusioni saranno commentati gli stimatori proposti, riassunti i risultati ottenuti e presentati i futuri scenari di ricerca.

In Appendice, inoltre, sono riportate in maniera dettagliata le dimostrazioni dei risultati presentati.

Capitolo 1

Aspetti introduttivi

Introduzione

L'obiettivo di questo capitolo è quello di richiamare i concetti alla base della teoria dei campioni e di illustrare tutti quegli argomenti che saranno propedeutici per i capitoli successivi.

Nel paragrafo 1.1 sarà introdotta la notazione adottata nel testo. Nel paragrafo 1.2 sarà illustrata brevemente la metodologia alla base degli stimatori lineari ed in particolar modo lo stimatore di Horvitz-Thompson (HT). Nel paragrafo 1.3 verrà chiarito il concetto di informazioni ausiliarie e, in quelli successivi, verranno illustrati lo stimatore di regressione generalizzata ($GREG$), paragrafo 1.4, e lo stimatore calibrato (CAL), paragrafo 1.5, che le impiegano per produrre stime più efficienti dello stimatore HT.

In questi ultimi due capitoli, sarà dedicata una sezione allo stimatore $GREG$ ottimo introdotto da Montanari (1987) e alla sua espressione equivalente in termini di stimatore CAL.

Infine, nel paragrafo 1.6 sarà presentato il contesto in cui il presente lavoro si colloca.

1.1 Quadro di riferimento

Si consideri la popolazione di riferimento

$$U = \{1, \cdots, k, \cdots, N\}$$

di numerosità N.
Da U si immagini di estrarre il campione

$$s = \{1, \cdots, k, \cdots, n\}$$

di numerosità fissata n attraverso un disegno campionario $(\mathcal{S}, p(\cdot))$ senza ripetizione in cui $\mathcal{S}$ è lo spazio dei campioni e $p(\cdot)$ una distribuzione di probabilità su $\mathcal{S}$ che soddisfa le condizioni:

$$\begin{aligned} &0 \leq p(s) \leq 1 \quad \forall\ s \in \mathcal{S} \\ &\sum_{s \in \mathcal{S}} p(s) = 1 \qquad\qquad . \end{aligned}$$

Definendo la variabile indicatrice per la presenza delle unità k nel campione,

$$I_k = \begin{cases} 1 & \text{se} \quad k \in s \\ 0 & \text{se} \quad k \notin s, \end{cases}$$

è generalmente possibile, a partire dal disegno campionario, determinare le probabilità di inclusione del primo ordine, ovvero la probabilità che l'unità k-ma sia contenuta nel campione. Tale quantità è data dalla somma delle probabilità di osservare tutti i campioni che contengono l'unità k, quindi

$$\pi_k = \Pr(k \in s) = \mathbb{E}[I_k] = \sum_{s \ni k} p(s)$$

con

$$\pi_k \geq 0 \quad \forall\ k \in U$$

e

$$\sum_{k \in U} \pi_k = n$$

in quanto la numerosità campionaria n è fissata.

Per tutte le unità k contenute nel campione s osservato si definisce, inoltre, la matrice diagonale

$$\boldsymbol{\Pi}_s = \operatorname{diag}\{\pi_k\}_{k \in s} = \begin{bmatrix} \pi_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \pi_k & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \pi_n \end{bmatrix} \tag{1.1}$$

che contiene le probabilità di inclusione del primo ordine.

Si consideri, inoltre, la probabilità di inclusione del secondo ordine, π_{kl}, che indica la probabilità che l'unità k e l'unità l siano contemporaneamente osservate nel campione,

$$\pi_{kl} = \Pr(k, l \in s) = \mathbb{E}[I_k \; I_l] = \sum_{s \ni k,l} p(s)$$

ottenuta, in maniera speculare a π_k, dalla somma delle probabilità di tutti i campioni che contengono le unità k e l contemporaneamente. Come caso particolare si ha $\pi_{kk} = \pi_k$ e, poiché la numerosità n del campione è fissata,

$$\sum_{k \in U} \pi_{kl} = n \; \pi_l.$$

Definite queste quantità, è possibile determinare la matrice del disegno, $\mathbf{\Delta}$, che è una matrice semi-definita positiva di dimensione $N \times N$ in cui il generico elemento è dato da

$$\Delta_{kl} = Cov(I_k \; I_l) = \begin{cases} \frac{1-\pi_k}{\pi_k} & \text{se} \quad k = l \\ \frac{\pi_{kl} - \pi_k \; \pi_l}{\pi_k \; \pi_l} & \text{se} \quad k \neq l \end{cases} \tag{1.2}$$

e con

$$\sum_{k \in U} \Delta_{kl} = 0$$

poiché la dimensione del campione è fissata.

La matrice del disegno può essere stimata dal campione. La sua stima, $\underline{\mathbf{\Delta}}$, è a sua volta una matrice semi-definita positiva di ordine $n \times n$ ed è ottenuta dividendo il generico elemento della matrice $\mathbf{\Delta}$ per la relativa probabilità di inclusione del secondo ordine, quindi

$$\underline{\mathbf{\Delta}} = \left(\frac{\Delta_{kl}}{\pi_{kl}} \right)_{k,l \in s} = \begin{bmatrix} \frac{\Delta_{11}}{\pi_1} & \cdots & \frac{\Delta_{1k}}{\pi_{1k}} & \cdots & \frac{\Delta_{1n}}{\pi_{1n}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\Delta_{k1}}{\pi_{k1}} & \cdots & \frac{\Delta_{kk}}{\pi_k} & \cdots & \frac{\Delta_{kn}}{\pi_{kn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\Delta_{n1}}{\pi_{n1}} & \cdots & \frac{\Delta_{nk}}{\pi_{nk}} & \cdots & \frac{\Delta_{nn}}{\pi_n} \end{bmatrix}$$

Si immagini di voler rilevare la variabile di interesse Y che assume valore y_k sulla generica unità k, con $k \in U$.

I valori y_k sono raccolti in un vettore $\mathbf{y} = (y_1 \cdots y_k \cdots y_N)^t$, ma sono noti solamente sulle unità del campione, $k \in s$. Definiamo, quindi, il vettore

$$\mathbf{y}_s = (y_1 \cdots y_k \cdots y_n)^t \tag{1.3}$$

il vettore che contiene i valori di y_k rilevati sulle unità $k \in s$. Il nostro obiettivo è quello di stimare il totale della variabile Y nella popolazione, dato da

$$t_Y = \sum_{k \in U} y_k.$$

In tutto il prosieguo della trattazione il parametro da stimare sarà il totale della variabile Y. Tale scelta non fa perdere di generalità ai risultati ottenuti, in quanto numerosi parametri possono essere espressi come combinazione lineare di totali (ad esempio la media e le proporzioni) e la loro varianza campionaria può essere calcolata in funzione delle stime delle varianze e covarianze campionarie degli stimatori di totali.

1.2 Lo stimatore di Horvitz-Thompson

Per stimare il totale della Y nella popolazione ricorriamo ad una classe fondamentale di stimatori, quella degli stimatori lineari. Gli stimatori lineari assumono la forma

$$\hat{t}_Y = a_{s0} + \sum_{k \in s} a_{sk}\ y_k,$$

dove a_{s0} è un coefficiente che dipende dal campione s, mentre a_{sk} sono pesi che vengono assegnanti alle unità campionarie e possono dipendere sia dalle unità che dal campione s.

Nel caso in cui il termine a_{s0} è nullo,

$$\hat{t}_Y = \sum_{k \in s} a_{sk}\ y_k$$

è detto lineare omogeneo (cfr. Cassel *e altri*, 1977).

Questi stimatori assumono praticamente la stessa forma del parametro nella popolazione, con l'unica differenza che nell'espressione del parametro nella popolazione tutte le unità hanno peso 1, mentre nello stimatore lineare le unità estratte k nel campione s hanno peso a_{sk}. I pesi a_{sk} possono, dunque, essere interpretati come il numero di unità nella popolazione rappresentate dall'unità k nel campione s.

Tra gli stimatori lineari, il più noto è lo stimatore di Horvitz-Thompson, HT (Horvitz e Thompson, 1952). É definito come

$$\hat{t}_{Y_{HT}} = \sum_{k \in s} d_k\ y_k. \tag{1.4}$$

Lo stimatore HT assegna a ciascuna unità estratta nel campione un peso $d_k = \pi_k^{-1}$, ovvero uguale all'inverso della probabilità di inclusione del primo ordine. Il peso d_k è anche chiamato peso base o peso da disegno.

La (1.4) può anche essere espressa in forma vettoriale

$$\hat{t}_{Y_{HT}} = \mathbf{y}_s{}^t \, \mathbf{d}_s \tag{1.5}$$

dove i pesi campionari sono raccolti nel vettore $\mathbf{d}_s = (d_{1s} \cdots d_{ks} \cdots d_{ns})^t$.

Lo stimatore HT è uno stimatore corretto del totale nella popolazione,

$$\mathbb{E}\left[\hat{t}_{Y_{HT}}\right] = t_Y,$$

e la sua varianza nella popolazione è

$$Var\left(\hat{t}_{Y_{HT}}\right) = \sum_{k=1}^{N} \sum_{l=1}^{N} \Delta_{kl} \; y_k \; y_l \tag{1.6}$$

con Δ_{kl} definito nella (1.2). Uno stimatore corretto della varianza nella popolazione è

$$var\left(\hat{t}_{Y_{HT}}\right) = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\Delta_{kl}}{\pi_{kl}} \; y_k \; y_l \tag{1.7}$$

La (1.6) e la (1.7) possono essere espresse in forma matriciale

$$\begin{aligned} Var\left(\hat{t}_{Y_{HT}}\right) &= \mathbf{y}^t \; \mathbf{\Delta} \; \mathbf{y} \\ var\left(\hat{t}_{Y_{HT}}\right) &= \mathbf{y}_s{}^t \; \underline{\mathbf{\Delta}} \; \mathbf{y}_s \end{aligned} \tag{1.8}$$

utilizzando la matrice del disegno, $\mathbf{\Delta}$, e la matrice del disegno stimata sul campione, $\underline{\mathbf{\Delta}}$, viste in precedenza.

1.3 Le informazioni ausiliarie

Nelle indagini campionarie, soprattutto in quelle su larga scala, oltre alla variabile Y vengono rilevate sulle unità del campione altre variabili, come ad esempio età, sesso, stato civile, etc. Queste variabili vengono generalmente rilevate per produrre delle stime del parametro di interesse per diversi aggregati e studiarne, ad esempio, la distribuzione a seconda dell'età o del sesso, ma possono essere impiegate anche per migliorare l'efficienza delle stime.

Ipotizziamo, dunque, che oltre alla variabile di interesse Y, nella popolazione sia disponibile un set di variabili ausiliarie $X_1 \cdots X_p \cdots X_P$.

I vettori $\mathbf{x}_k = (x_{k1} \cdots x_{kp} \cdots x_{kP})^t$, relativi a ciascuna unità della popolazione, sono raccolti nella matrice della popolazione $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_k \cdots \mathbf{x}_N)^t$.

Su ciascuna unità del campione, $k \in s$, osserviamo i valori $(y_k \ \mathbf{x}_k)$. Dunque, oltre al vettore $\mathbf{y}_s$, già definito nella (1.3), definiamo la matrice campionaria $\mathbf{X}_s = (\mathbf{x}_1 \cdots \mathbf{x}_k \cdots \mathbf{x}_n)^t$.

Assumiamo che i totali nella popolazione delle variabili ausiliarie siano conosciuti da fonte amministrativa o fonte censuaria, quindi non affetti da errore campionario. Questi totali, chiamati anche totali di controllo o totali noti possono essere utilizzati per costruire stimatori più efficienti dello stimatore HT.

Definiamo, dunque, il vettore dei totali di controllo

$$\mathbf{t}_{\mathbf{X}} = \sum_{k \in U} \mathbf{x}_k = \left(t_{x_1} \cdots t_{x_p} \cdots t_{x_P}\right)^t, \tag{1.9}$$

che dunque avranno, per definizione, matrice di varianza e covarianza nulla,

$$Var\left(\mathbf{t}_{\mathbf{X}}\right) = \underline{\mathbf{0}}.$$

Relativamente all'esempio fatto in precedenza per le variabili ausiliarie, i totali di controllo potrebbero essere ricavati dalla distribuzione per età, sesso e stato civile fornita dall'anagrafe.

1.4 Lo stimatore di regressione generalizzata

Una delle più importanti procedure che usa informazioni a livello di popolazione per costruire stimatori efficienti è la procedura di regressione. L'uso della regressione nelle indagini campionarie risale già agli anni '40, ma è tra gli anni '70 ed '80 che si afferma (cfr. Fuller, 2002, per un'ampia rassegna della letteratura sul tema). In questo periodo, infatti, risalgono gli studi sulla natura generale dello stimatore di regressione nelle indagini campionarie e sul grado con il quale l'approccio predittivo relativo al modello può essere conciliato con la prospettiva del disegno campionario.

Il lavoro di Cassel *e altri* (1979) costituisce, in questo contesto, un punto di riferimento importantissimo che fornisce la base per un approccio cosiddetto *model-assisted*, e grazie anche al loro contributo la procedura di regressione è dagli anni '90 largamente diffusa nel processo di produzione di stime da indagini campionarie.

Cassel *e altri* (1976), nel caso in cui sono disponibili informazioni ausiliarie, propongono di costruire lo stimatore di regressione generalizzato ($GREG$) seguendo sia i principi del modello che del disegno campionario. Se sono disponibili variabi-

li ausiliarie correlate con la variabile di interesse può essere, infatti, costruito uno stimatore più preciso dello stimatore HT[1].

Lo stimatore $GREG$ assume implicitamente l'esistenza di un modello di regressione lineare di superpopolazione,

$$\xi : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

che spiega la relazione tra la variabile di interesse ed il set di variabili ausiliarie, dove $\mathbf{e} = (e_1 \cdots e_k \cdots e_N)^t$ è il vettore dei residui, in cui i generici elementi sono variabili aleatorie incorrelate con valore atteso rispetto al modello nullo e varianza, sempre rispetto al modello, σ_k^2. In formule

- $\mathbb{E}_\xi [\mathbf{e}] = \mathbf{0}$
- $Var_\xi (\mathbf{e}) = \sigma_k^2 \, \mathbf{I} = \frac{\sigma^2}{q_k}\mathbf{I}$

dove il termine q_k è un valore noto e positivo che serve ad attribuire maggior o minor importanza all'unità k-ma, generalmente $q_k = 1$, e $\mathbf{I}$ è la matrice identità di ordine N (cfr. Bethlehem e Keller, 1987; Cassel *e altri*, 1979; Isaki e Fuller, 1982; Luery, 1986; Särndal, 1980; Särndal *e altri*, 1989; Wright, 1983).
Il termine

$$\boldsymbol{\beta} = (\beta_1 \cdots \beta_p \cdots \beta_P)^t$$

è il vettore dei coefficienti di regressione incogniti delle P variabili ausiliarie.

Dalla teoria del modello di regressione generale si ha che, se la matrice $\left(\mathbf{X}^t \, \mathbf{X}\right)$ è invertibile, l'espressione dei coefficienti di regressione $\boldsymbol{\beta}$ che minimizza l'errore quadratico medio, rispetto al modello ξ, è

$$\begin{aligned} \boldsymbol{\beta} = (\beta_1 \cdots \beta_p \cdots \beta_P)^t &= \left(\sum_{k \in U} \frac{q_k \, \mathbf{x}_k \, \mathbf{x}_k^t}{\sigma^2} \right)^{-1} \left(\sum_{k \in U} \frac{q_k \, \mathbf{x}_k \, y_k}{\sigma^2} \right) \\ &= \left(\mathbf{X}^t \, \mathbf{X} \right)^{-1} \mathbf{X}^t \, \mathbf{y} \end{aligned}$$

ed un suo stimatore basato sui dati campionari raccolti è dato da

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1 \cdots \hat{\beta}_p \cdots \hat{\beta}_P \right)^t &= \left(\sum_{k \in s} \frac{q_k \, \mathbf{x}_k \, \mathbf{x}_k^t}{\pi_k \, \sigma^2} \right)^{-1} \left(\sum_{k \in s} \frac{q_k \, \mathbf{x}_k \, y_k}{\pi_k \, \sigma^2} \right) \\ &= \left(\mathbf{X}_s{}^t \, \boldsymbol{\Pi}_s^{-1} \, \mathbf{X}_s \right)^{-1} \mathbf{X}_s{}^t \, \boldsymbol{\Pi}_s^{-1} \, \mathbf{y}_s \end{aligned} \tag{1.10}$$

[1] "[...] *if the auxiliary variables are correlated with the targer variables, an estimator that is more precise the Horvitz-Thompson estimator can be constructed*" (Bethlehem e Keller, 1987, p. 143).

dove rispetto alla definizione data nella (1.1), $\mathbf{\Pi} = \text{diag}\left(q_k^{-1}\ \pi_k\right)_{k\in s}$.
Lo stimatore $\hat{\boldsymbol{\beta}}$ nella (1.10) è ottenuto calcolando i coefficienti di regressione pesati con pesi uguali all'inverso della probabilità di inclusione del primo ordine (i.e. uguali ai pesi da disegno).

Sfruttando la relazione data dal modello ξ e ricorrendo alla (1.10) è possibile calcolare il totale della Y attraverso l'espressione dello stimatore $GREG$ così definito

$$\hat{t}_{Y_{GREG}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}} \tag{1.11}$$

dove $\hat{t}_{Y_{HT}}$ è la stima del totale di Y ottenuta con lo stimatore HT nella (1.4), $\hat{\mathbf{t}}_{\mathbf{X}_{HT}}$ sono le stime di HT dei totali delle variabili ausiliarie e $\mathbf{t_X}$ il vettore dei totali di controllo definito nella (1.9).

Lo stimatore $GREG$, definito come nella (1.11), può anche essere visto come lo stimatore alle differenze generalizzato con varianza minima rispetto al disegno, ottenuto stimando il vettore incognito $\boldsymbol{\beta}$ con il metodo dei minimi quadrati generalizzati (cfr. Cassel *e altri*, 1976; Cicchitelli *e altri*, 1992).

Un'espressione alternativa e molto utile dello stimatore $GREG$ è

$$\hat{t}_{Y_{GREG}} = \sum_{k\in s} d_k\ \gamma_{ks}\ y_k,$$

dove compare il fattore γ_{ks}, anche chiamato coefficiente di correzione dei pesi base in quanto viene moltiplicato per d_k, che è pari a

$$\gamma_{ks} = 1 + \mathbf{x}_k^t \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \left(\mathbf{X}^t\ \mathbf{\Pi}^{-1}\ \mathbf{X}\right)^{-1}. \tag{1.12}$$

Da questa scrittura alternativa dello stimatore $GREG$ appare evidente come questo rientri nella categoria degli stimatori lineari omogenei.

La sua espressione della varianza è funzione dei residui del modello, $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$,

$$Var\left(\hat{t}_{Y_{GREG}}\right) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\ e_k\ e_l \tag{1.13}$$

e uno stimatore della varianza è dato da

$$var\left(\hat{t}_{Y_{GREG}}\right) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_k\ \hat{e}_k\right)\left(\gamma_l\ \hat{e}_l\right), \tag{1.14}$$

dove compaiono i residui stimati sulle unità del campione s, $\hat{e}_k = y_k - \mathbf{x}_k^t\hat{\boldsymbol{\beta}}$ (Särndal *e altri*, 1989).

Nel caso in cui nella (1.14) si considera $\gamma_{ks} = 1$ si ottiene lo stimatore della varianza di $\hat{t}_{Y_{GREG}}$ derivato con il metodo classico della linearizzazione di Taylor.

L'espressione così ottenuta, che si può scrivere come

$$var\left(\hat{t}_{Y_{GREG}}\right) \approx \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\,\hat{e}_k\,\hat{e}_l \tag{1.15}$$

$$= \hat{\mathbf{e}}_s{}^t\,\underline{\boldsymbol{\Delta}}\,\hat{\mathbf{e}}_s \tag{1.16}$$

rappresenta un'approssimazione della (1.14) che comporta una leggera sottostima soprattutto per campioni di piccole dimensioni (Estevao *e altri*, 1995, pp. 184-185). Nella (1.16) compare il vettore dei residui campionari $\hat{\mathbf{e}}_s = \left(\hat{e}_k = y_k - \mathbf{x}_k^t\hat{\boldsymbol{\beta}}\right)_{k\in s}$.

Lo stimatore $GREG$ gode di importanti proprietà, sia rispetto al modello (ξ) che rispetto al disegno (d), infatti è

- non-distorto rispetto al modello

$$\mathbb{E}_\xi\left[\mathbf{y}\right] = \mathbf{X}\boldsymbol{\beta}$$

- asintoticamente non-distorto rispetto al disegno (ADU, *asymptotical design unbiased*)

$$\mathbb{E}_d\left[\hat{t}_{Y_{GREG}}\right] \asymp t_Y$$

- asintoticamente consistente rispetto al disegno

$$MSE_d\left(\hat{t}_{Y_{GREG}}\right) \asymp 0.$$

La sua distorsione è dell'ordine $n^{-1/2}$, quindi decresce rapidamente al crescere della numerosità campionaria n. Isaki e Fuller (1982) e Robinson e Särndal (1983) hanno dimostrato, inoltre, che le proprietà di non-distorsione e di consistenza rispetto al disegno sono verificate anche quando il modello non è corretto.

Un'altra proprietà che soddisfa lo stimatore $GREG$, e che sarà richiamata più volte in seguito, è la proprietà di coerenza esterna. Stimando sul campione s i totali delle variabili ausiliarie con lo stimatore $GREG$ si ottengono esattamente i totali della popolazione noti da fonte amministrativa, $\mathbf{t_X}$. Infatti, dalla (1.11)

$$\begin{aligned}\hat{\mathbf{t}}_{\mathbf{X}_{GREG}} &= \hat{\mathbf{t}}_{\mathbf{X}_{HT}} + \hat{\boldsymbol{\beta}}^t\left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\\ &= \hat{\mathbf{t}}_{\mathbf{X}_{HT}} + \mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ &= \mathbf{t_X},\end{aligned}$$

in quanto $\boldsymbol{\beta}$ nella (1.10) si riduce ad una matrice identità di ordine P. Lo stesso risultato si ottiene considerando il coefficiente di correzione dei pesi da disegno, γ_{ks} nella (1.12).

1.4.1 Lo stimatore di regressione generalizzata ottimo

Montanari (1987) individua l'espressione dello stimatore $GREG$ con varianza minima nella classe degli stimatori di regressione generalizzata che utilizzano lo stesso set di variabili ausiliarie.

Nel caso di un disegno campionario con n fissato, la varianza rispetto al disegno dello stimatore $GREG$, così come è espresso nella (1.11), è data da

$$Var\left(\hat{t}_{Y_{GREG}}\right) = \mathbf{y}^t \ \boldsymbol{\Delta} \ \mathbf{y} + \boldsymbol{\beta}^t \mathbf{X}^t \ \boldsymbol{\Delta} \ \mathbf{X}\boldsymbol{\beta} - 2 \ \mathbf{y}^t \ \boldsymbol{\Delta} \ \mathbf{X}\boldsymbol{\beta}, \tag{1.17}$$

in cui $\boldsymbol{\Delta}$ è la matrice del disegno.

Assumendo che $\boldsymbol{\beta}$ non sia una variabile aleatoria ma sia costante, è possibile minimizzare l'espressione della varianza da disegno di $\hat{t}_{Y_{GREG}}$ rispetto al vettore dei coefficienti di regressione. Si ottiene, infatti, che il vettore dei coefficienti di regressione,

$$\begin{aligned} \boldsymbol{\beta}_o &= \left(\sum_{k\in s}\sum_{l\in s} \Delta_{kl} \ \mathbf{x}_k \ \mathbf{x}_l^t\right)^{-1} \left(\sum_{k\in s}\sum_{l\in s} \Delta_{kl} \ \mathbf{x}_k \ y_l\right) \\ &= \left(\mathbf{X}^t \ \boldsymbol{\Delta} \ \mathbf{X}\right)^{-1} \left(\mathbf{X}^t \ \boldsymbol{\Delta} \ \mathbf{y}\right), \end{aligned} \tag{1.18}$$

consente di ottenere la varianza minima tra la classe di stimatori $GREG$ che utilizzano le stesse variabili ausiliarie.

Lo stimatore proposto da Montanari, non è realmente ottimo, in quanto in pratica $\boldsymbol{\beta}_o$ deve essere stimato. Un possibile stimatore di $\boldsymbol{\beta}_o$ è

$$\begin{aligned} \hat{\boldsymbol{\beta}}_o &= \left(\sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \ \mathbf{x}_k \ \mathbf{x}_l^t\right)^{-1} \left(\sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \ \mathbf{x}_k \ y_l\right) \\ &= \left({\mathbf{X}_s}^t \ \underline{\boldsymbol{\Delta}} \ \mathbf{X}_s\right)^{-1} \left({\mathbf{X}_s}^t \ \underline{\boldsymbol{\Delta}} \ \mathbf{y}_s\right), \end{aligned}$$

dove $\underline{\boldsymbol{\Delta}}$ è la matrice del disegno stimata sul campione e, quindi,

$$\hat{t}_{Y_{GREG_o}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}_o. \tag{1.19}$$

Lo stimatore $GREG_o$, così come lo stimatore $GREG$, soddisfa la proprietà di coerenza esterna. Tuttavia, $\hat{\boldsymbol{\beta}}_o$ è una variabile casuale e l'ottimalità è raggiunta solamente in via approssimata, in quanto $\hat{\boldsymbol{\beta}}_o$ converge in probabilità a $\boldsymbol{\beta}_o$.

Questo, comunque, consente di determinare la riduzione della varianza da disegno che si ottiene passando dallo stimatore $GREG_o$ allo stimatore HT, che è pari a

$$deft\left(\hat{t}_{Y_{GREG}}\right) = \frac{var\left(\hat{t}_{Y_{GREG}}\right)}{var\left(\hat{t}_{Y_{HT}}\right)} = {\mathbf{y}_s}^t \ \boldsymbol{\Delta}_s \ \mathbf{X}_s \left({\mathbf{X}_s}^t \ \boldsymbol{\Delta}_s \ \mathbf{X}_s\right)^{-1} {\mathbf{X}_s}^t \ \boldsymbol{\Delta}_s \ \mathbf{y}_s$$

ed è ottenuta facendo il rapporto tra la (1.17) e la (1.8). Questa quantità è una forma quadratica semi-definita positiva, quindi vuol dire che è sempre possibile ottenere un guadagno di efficienza, ovviamente in presenza di un campione con un'adeguata numerosità campionaria. Questo risultato è indipendente dalla validità del modello e, in questo contesto, l'uso del modello si riduce alla scelta delle variabili rilevanti per predire i valori della variabile di interesse.

Tuttavia, questo metodo di stima non è facilmente applicabile, perché la stima dei coefficienti regressione può essere particolarmente complicata in quanto coinvolge la matrice $\boldsymbol{\Delta}$ che dipende dalle probabilità di inclusione del secondo ordine. Un altro problema dello stimatore $GREG_o$ è rappresentato dalla possibilità che le stime prodotte siano instabili, caso che si presenta quando il numero dei gradi di libertà, $n - P - 1$, è piccolo (Rao, 1994; Montanari, 1988).

1.5 Lo stimatore calibrato

Un altro stimatore che fornisce un modo sistematico per inglobare variabili ausiliarie nella procedura di stime è lo stimatore calibrato (CAL, *calibrated estimator*), in alcuni casi chiamato anche stimatore di ponderazione vincolata.

Lo stimatore CAL definisce una classe di stimatori che godono della proprietà della coerenza esterna che comprende, dunque, anche lo stimatore $GREG$. Da quando è stato proposto da Deville e Särndal (1992) è stato al centro di un'intensa produzione scientifica ed ha fornito un importante strumento per la produzione di indagini su larga scala fornendo, inoltre, la possibilità di fronteggiare meglio il problema della mancata risposta (cfr. Särndal e Lunström, 2005; Kim *e altri*, 2007; Kott, 2006) e della sotto-copertura. È stato, infatti, adottato da diversi Istituti nazionali di statistica che hanno costruito dei software *ad hoc*.

Särndal (2007, p. 99), proprio per questo grande successo, più che di stimatore calibrato, parla di approccio di calibrazione e lo definisce come

- il calcolo di pesi che incorporano specifiche informazioni ausiliarie e che sono vincolati da equazioni di calibrazione;

- l'uso di pesi per calcolare stime lineari di totali o di altri parametri della popolazione;

- una metodologia per ottenere stime approssimativamente non-distorte in caso di assenza di mancata risposta e di altri errori non campionari.

Seppur molto simile come finalità allo stimatore $GREG$, lo stimatore CAL si basa su un'idea completamente differente. Ovvero, se si conoscono nella popolazione i totali di alcune variabili ausiliarie fortemente correlate con la variabile di interesse e si è in grado di costruire un sistema di pesi che conduce a buone stime per le variabili ausiliarie (cioè è soddisfatta la proprietà di coerenza esterna), allora si hanno buone garanzie che lo stesso sistema di pesi, applicato alla variabile di interesse, condurrà a buone stime[2].

Lo stimatore CAL può essere scritto come uno stimatore lineare omogeneo

$$\hat{t}_{Y_{CAL}} = \sum_{k \in s} w_k \; y_k = \sum_{k \in s} d_k \; \gamma_{ks} \; y_k \tag{1.20}$$

$$= \mathbf{w}_s{}^t \; \mathbf{y}_s, \tag{1.21}$$

dove $\mathbf{w}_s$ è il vettore dei pesi finali, in cui il generico elemento $w_k = \gamma_{ks} \; d_k$ per $k \in s$. Il sistema di pesi w_k è determinato risolvendo il problema di minimo vincolato

$$\begin{cases} \min_{w_k} \left\{ \sum_{k \in s} G_k \left(w_k - d_k \right) / q_k \right\} \\ \\ \sum_{k \in s} w_k \; \mathbf{x}_k = \mathbf{t}_{\mathbf{X}} \end{cases} \tag{1.22}$$

in cui la funzione obiettivo da minimizzare rispetto a w_k,

$$\sum_{k \in s} G_k \left(w_k - d_k \right) / q_k,$$

garantisce che il sistema di pesi w_k sia il più vicino possibile, in media rispetto alla metrica $G(\cdot)$, ai pesi da disegno d_k con q_k, generalmente $q_k = 1$, quantità nota e positiva che attribuisce più o meno importanza all'unità k. In questo modo è garantita la proprietà di non-distorsione rispetto al disegno, almeno in via asintotica, in quanto il sistema di pesi ottenuto è il più vicino possibile a quello adottato dallo stimatore HT che è non-distorto rispetto al disegno.

È bene precisare che, in pratica, i pesi d_k utilizzati come base di partenza per il problema di minimo vincolato sono già aggiustati per mancata risposta (Deville *e altri*, 1993) e che il vettore $\mathbf{x}_k$ di variabili ausiliarie, nelle indagini sulle famiglie, è costruito nello stesso modo per tutti gli individui della stessa famiglia, in modo da garantire che tutti gli individui della stessa famiglia abbiano lo stesso peso w_k (Lemaître e Dufour, 1987).

[2] " [...] *a strong correlation between the auxiliary variables and the study variable means that the weights that perform well for the auxiliary variables also should perform well for the interest variables*" (Deville e Särndal, 1992, p. 376).

Tabella 1.1: Funzioni di pseudo-distanza $G(\cdot)$, relativo fattore di correzione dei pesi base e intervallo del sistema di pesi finali.

	Funzione	$G_k\,(w_k - d_k)$	γ_{ks}	Range w_k
a.	Lineare	$\frac{(w_k - d_k)^2}{2d_k}$	$1 + \mathbf{x}_k^t\ \boldsymbol{\lambda}$	$-\infty \leq w_k \leq +\infty$
b.	Logaritmica	$w_k\ ln\left(\frac{w_k}{d_k}\right) - w_k + d_k$	$\exp\left(\mathbf{x}_k^t\ \boldsymbol{\lambda}\right)$	$0 \leq w_k \leq +\infty$
c.	Chi-quadrato	$\frac{1}{2\ w_k}\left(\frac{w_k}{d_k} - 1\right)^2$	$\left(1 - 2\ \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)^{-1/2}$	$0 \leq w_k \leq +\infty$
d.	Minima entropia	$-d_k\ ln\left(\frac{w_k}{d_k}\right) + w_k - d_k$	$\left(1 - \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)^{-1}$	$0 \leq w_k \leq +\infty$
e.	Hellinger	$2\ d_k\left(\sqrt{\frac{w_k}{d_k}} - 1\right)^2$	$\left(1 - \frac{1}{2}\ \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)^{-2}$	$0 \leq w_k \leq +\infty$
f.	Logaritmica Troncata	$\left(\frac{w_k}{d_k} - L\right) ln\left(\frac{\frac{w_k}{d_k} - L}{1 - L}\right) + \left(U - \frac{w_k}{d_k}\right) ln\left(\frac{U - \frac{w_k}{d_k}}{U - 1}\right)$	$\frac{L(U-1) + U(1-L)\ \exp\left(\frac{U-L}{(U-1)(1-L)}\ \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)}{(U-1) + (1-L)\ \exp\left(\frac{U-L}{(U-1)(1-L)}\ \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)}$	$L\ \leq w_k \leq\ U$

Il vincolo imposto,

$$\sum_{k \in s} w_k \ \mathbf{x}_k = \mathbf{t}_{\mathbf{X}},$$

fa sì che lo stimatore CAL soddisfi la condizione di coerenza esterna e che quindi il sistema di pesi sia calibrato sui totali delle variabili ausiliarie note da fonti amministrative (Deville e Särndal, 1992; Singh e Mohl, 1996; Stukel *e altri*, 1996).

La funzione di pseudo-distanza $G(\cdot)$ deve soddisfare alcune proprietà di regolarità, ovvero

- la distanza tra w_k e d_k misurata con G_k deve sempre essere non negativa,

$$w_k \mapsto G_k(w_k - d_k) \geq 0;$$

- deve essere strettamente convessa;

- deve essere continuamente differenziabile rispetto a w_k;

- la distanza tra un punto e se stesso misurata con G_k deve essere nulla,

$$G_k(d_k - d_k) = 0.$$

In Deville e Särndal (1992) ed in Singh e Mohl (1996) sono illustrate diverse funzioni di pseudo-distanza che producono diversi sistemi di pesi (cfr. Tabella 1.1).

La scelta della funzione di pseudo-distanza conduce a sistemi di pesi diversi e, quindi, a stime leggermente differenti in media per grandi campioni. La funzione di distanza lineare può portare alla determinazione di pesi finali negativi o particolarmente grandi, che sono di difficile interpretazione e difficile gestione nelle indagini su larga scala. Questo può avvenire quando si estrae un campione sbilanciato rispetto alle variabili ausiliarie o quando il numero di vincoli è particolarmente grande in relazione alla dimensione del campione. La funzione logaritmica, quella del chi-quadrato e quella di Hellinger prevengono la determinazione di pesi negativi, ma non sono in grado di prevenire pesi particolarmente grandi.

La funzione di pseudo-distanza logaritmica troncata, invece, consente di ottenere dei pesi che cadono all'interno di un intervallo prestabilito, $[L, U]$ e, per questo motivo, è quella che viene impiegata nelle indagini su larga scala da parte degli Istituti nazionali di statistica. Gli estremi dell'intervallo in cui cadono i pesi base sono definiti dal ricercatore. In generale, per preservare la proprietà di non-distorsione rispetto al disegno si determinano in modo tale da contenere la variazione del singolo peso finale rispetto al relativo peso da disegno, quindi

$$L = l \ d_k$$

$$U = u\ d_k.$$

Questa soluzione non garantisce sempre l'esistenza della soluzione al problema di minimo vincolato. In pratica, quindi, si parte con l'attribuire un valore prossimo a 0 per l ed a 1 per u e si procede allargando di volta in volta leggermente l'intervallo fin quando il problema di minimo vincolato non ha soluzione.

Alternative a questa soluzione euristica del problema sono fornite da Théberge (1999, 2000) che individua un metodo che garantisce sempre l'esistenza della soluzione per la (1.22) e da Bardsley e Chambers (1984); Chambers (1996); Rao e Singh (1997), che, invece, seguono un approccio *model-based*. In quest'ottica possono essere letti anche i risultati ottenuti da Guggemos e Tillé (2010).

Tutte le funzioni riportate in Tabella 1.1 determinano un fattore di correzione dei pesi base lineare rispetto a $\mathbf{x}_k$

$$\gamma_{ks} = \left(1 + a\ \mathbf{x}_k^t\ \boldsymbol{\lambda}\right)^{1/a}$$

con a che assume valori diversi a seconda delle funzioni di distanza e con $\boldsymbol{\lambda}$ vettore dei moltiplicatori di Lagrange che risolvono il sistema nella (1.22) se esiste l'inversa di $\left(\mathbf{X}_s{}^t\ \boldsymbol{\Pi}_s^{-1}\ \mathbf{X}_s\right)$. Rispetto a tutte, però, particolarmente importante è la funzione di distanza del chi-quadrato.

Come dimostrato da Deville e Särndal (1992), minimizzando la distanza dei pesi w_k rispetto ai pesi d_k con questa metrica, lo stimatore calibrato è equivalente allo stimatore $GREG$ nella (1.11).

Infatti, con la pseudo-distanza del chi-quadrato, si ha che

$$\boldsymbol{\lambda} = \left(\mathbf{X}_s{}^t\ \boldsymbol{\Pi}_s^{-1}\ \mathbf{X}_s\right)^{-1} \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)$$

e γ_{ks} è uguale alla (1.12).

Il vettore dei pesi finali che ne deriva[3] è uguale a

$$\mathbf{w}_s = \mathbf{d}_s + \boldsymbol{\Pi}_s^{-1}\ \mathbf{X}_s \left(\mathbf{X}_s{}^t\ \boldsymbol{\Pi}_s^{-1}\ \mathbf{X}_s\right)^{-1} \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \tag{1.23}$$

in cui il generico elemento $k \in s$ del vettore $\mathbf{w}_s$ è

$$\begin{aligned} w_k &= d_k\ \gamma_{ks} = \gamma_{ks} \\ &= d_k \left(1 + \mathbf{x}_k^t \left(\mathbf{X}_s{}^t\ \boldsymbol{\Pi}_s^{-1}\ \mathbf{X}_s\right)^{-1} \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\right). \end{aligned}$$

[3] Si veda dimostrazione in Appendice, A1.

Se si sostituisce la (1.23) nella (1.21) e si considerano la (1.5) e la (1.10) si ottiene esattamente l'espressione dello stimatore $GREG$ già presentata nella (1.11), infatti

$$\begin{aligned}\hat{t}_{Y_{CAL}} &= \mathbf{w}_s{}^t\, \mathbf{y}_s \\ &= \mathbf{d}_s{}^t\, \mathbf{y}_s + \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \left(\mathbf{X}_s{}^t\, \mathbf{\Pi}_s^{-1}\, \mathbf{X}_s\right)^{-1} \mathbf{X}_s\, \mathbf{\Pi}_s^{-1} \mathbf{y}_s \\ &= \hat{t}_{Y_{HT}} + \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}.\end{aligned}$$

Questo risultato è verificato al crescere della dimensione campionaria anche per le altre funzione di pseudo-distanza presentate ed è particolarmente importante in quanto consente di determinare un'espressione in forma chiusa della varianza dello stimatore CAL.

L'espressione della varianza asintotica dello stimatore CAL è data da:

$$AVar\left(\hat{t}_{Y_{CAL}}\right) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\; e_k\; e_l$$

e uno stimatore della varianza asintotica è dato da

$$Avar\left(\hat{t}_{Y_{CAL}}\right) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_k\; \hat{e}_k\right)\left(\gamma_l\; \hat{e}_l\right), \tag{1.24}$$

che sono equivalenti alle espressioni (1.13) e (1.14) già presentate per lo stimatore $GREG$. Anche per lo stimatore CAL valgono, dunque, le espressioni approssimate dello stimatore della varianza presentate nella (1.15) e nella (1.16).

Riassumendo quanto detto, lo stimatore CAL

- è asintoticamente non-distorto rispetto al disegno (ADU, *asymptotical design unbiased*)
$$\mathbb{E}_d\left[\hat{t}_{Y_{CAL}}\right] \asymp t_Y$$
- è asintoticamente consistente rispetto al disegno
$$MSE_d\left(\hat{t}_{Y_{CAL}}\right) \asymp 0.$$
- soddisfa la proprietà di coerenza esterna[4],

$$\begin{aligned}\hat{\mathbf{t}}_{\mathbf{X}_{CAL}} &= \mathbf{X}_s^t\, \mathbf{w}_s \\ &= \mathbf{y}_s{}^t\, \mathbf{d}_s + \mathbf{X}_s{}^t\, \mathbf{\Pi}_s^{-1}\, \mathbf{X}_s \left(\mathbf{X}_s^t\, \mathbf{\Pi}_s^{-1}\, \mathbf{X}_s\right)^{-1} \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\ &= \hat{\mathbf{t}}_{\mathbf{X}_{HT}} + \mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}} \\ &= \mathbf{t}_X.\end{aligned}$$

[4]Questa proprietà discende direttamente dalla definizione del problema di minimo vincolato nella (1.22), in quanto questa condizione rappresenta il vincolo imposto.

1.5.1 Lo stimatore "calibrato" ottimo

Nel paragrafo precedente abbiamo illustrato gli stretti legami che intercorrono tra lo stimatore $GREG$ e lo stimatore CAL. Lo stesso stimatore $GREG$ ottimo di Montanati, presentato nel paragrafo 1.4.1, può essere scritto nella forma dello stimatore CAL.

È sufficiente, infatti, risolvere il problema di minimo vincolato

$$\begin{cases} \min_{w_k} \left\{ \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}\ q_k} \left(w_k - d_k\right) \left(w_l - d_l\right) \right\} \\ \\ \sum_{k \in s} w_k\ \mathbf{x}_k = \mathbf{t_X} \end{cases} \tag{1.25}$$

Il sistema di pesi finali che viene individuato[5] è, quindi

$$\mathbf{w}_{GREG_o} = \mathbf{d}_s + \boldsymbol{\Delta}_s\ \mathbf{X}_s \left(\mathbf{X}_s t\ \boldsymbol{\Delta}_s\ \mathbf{X}_s\right)^{-1} \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \tag{1.26}$$

Deville (2000). Anche lo stimatore così definito soddisfa la proprietà di coerenza esterna.

1.6 Le informazioni ausiliarie da fonti campionarie

Quanto visto finora è valido nel caso in cui sulle unità del campione s viene rilevato un set di variabili ausiliarie $X_1 \cdots X_p \cdots X_P$ di cui si conoscono i totali

$$\mathbf{t_X} = \sum_{k \in U} \mathbf{x}_k = \left(t_{x_1} \cdots t_{x_p} \cdots t_{x_P}\right)^t,$$

da fonti non affette da errore campionario per cui, come già detto nel paragrafo 1.1, per definizione,

$$Var\left(\mathbf{t_X}\right) = \underline{\mathbf{0}}.$$

Ipotizziamo di aver rilevato sulle unità del campione s anche un secondo set di variabili ausiliarie $Z_1 \cdots Z_m \cdots Z_M$, di cui non conosciamo i totali nella popolazione,

$$\mathbf{t_Z} = \sum_{k \in U} \mathbf{z}_k = \left(t_{z_1} \cdots t_{z_m} \cdots t_{z_M}\right)^t,$$

ma per cui abbiamo delle stime ottenute attraverso un'altra indagine campionaria svolta su un campione che indicheremo con s_1. Da qui in seguito, indicheremo con

[5] Si veda dimostrazione in Appendice, A2

s_2 il campione dell'indagine in corso su cui vogliamo stimare il totale della Y e con s_1 il campione dell'indagine su cui sono stati stimati i totali delle variabili ausiliarie Z. In generale, quando sarà utilizzato il pedice 1 ci riferiremo a quantità relative all'indagine svolta su s_1 e, in maniera speculare, il pedice 2 indicherà quantità relative all'indagine svolta su s_2.

I totali stimati su s_1 saranno raccolti in un vettore di stime,

$$\widetilde{\mathbf{t}}_{\mathbf{Z}_1} = \left(\tilde{t}_{z_1} \cdots \tilde{t}_{z_m} \cdots \tilde{t}_{z_M}\right)^t,$$

che chiameremo stime di controllo o totali di controllo campionari. Per questo vettore di totali non possiamo assumere come fatto per $\mathbf{t_X}$ che la matrice di varianze e covarianze sia nulla, ma si avrà ovviamente che

$$Var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \neq \underline{\mathbf{0}}.$$

Come ampiamente illustrato nell'introduzione, questa situazione non è così rara e in un futuro prossimo l'esigenza di confrontarsi con informazioni ausiliarie con totali di controllo campionari sarà sempre maggiore.

Gli stimatori illustrati nei paragrafi precedenti sono ampiamente affidabili e le loro proprietà ampiamente testate quando si dispone di informazioni ausiliarie con totali di controllo non-campionari, ma devono essere adattati e testati nel caso in cui si voglia tener conto anche di informazioni ausiliarie con totali di controllo campionari.

Due sono le principali questioni che sorgono. La prima è relativa all'impatto che l'uso di totali di controllo affetti da errori campionari, al posto di totali noti, può avere sull'errore della stima della variabile di interesse. La seconda, invece, è relativa all'uso ottimale che può essere fatto di queste informazioni, condizionatamente alla natura campionaria del vettore dei totali. Riuscire a valutare in maniera adeguata questo permetterà, infatti, di decidere in maniera consapevole e appropriata se è conveniente utilizzare queste informazioni oppure se il loro impatto sull'errore finale è eccessivo ed è preferibile rinunciarvici.

Abbiamo, dunque, definito due differenti approcci che sottostano a due proprietà che possono essere desiderabili in questo contesto: la coerenza esterna e l'ottimalità. Il capitolo 2 sarà dedicato alla ricerca della coerenza esterna, mentre il capitolo 3 sarà dedicato alla ricerca dell'ottimalità.

Capitolo 2

Coerenza Esterna

Introduzione

In questo capitolo presenteremo la metodologia sviluppata nel caso in cui si richiede il rispetto della condizione di coerenza esterna con totali di controllo campionari, così come avviene per i totali di controllo non-campionari con lo stimatore $GREG$ e lo stimatore CAL.

I risultati che saranno presentati, innanzitutto, consentono di rispondere in maniera semplice e diretta ad un'esigenza particolarmente sentita dagli Istituti nazionali di statistica, ovvero, quella di costruire un sistema di indagini coerenti ed integrate. Infatti, nello stesso periodo sulla stessa popolazione sono svolte diverse indagini che molto spesso producono stime relative agli stessi aggregati. Le stime ottenute da due o più indagini differenti non necessariamente sono uguali e questo comporta un problema sul piano della comunicazione dell'informazione statistica da parte dell'Istituto.

Il secondo motivo, molto più importante da un punto di vista metodologico, è quello della valutazione dell'impatto che le informazioni ausiliarie con totali noti campionari hanno sull'errore della stima del parametro di interesse. Una prassi particolarmente diffusa, ma completamente errata, è quella di utilizzare le formule della varianza tradizionale, viste nel capitolo 1, come se i totali di controllo fossero noti senza errore. Infatti, Berger *e altri* (2009) e Dever e Valliant (2010) hanno dimostrato che con le espressioni della varianza tradizionale si sottostima notevolmente l'errore campionario.

In questo contesto un contributo particolarmente importante è quello fornito dall'Istituto nazionale di statistica olandese (CBS, Central Boureau of Statistique,

Statistics Netherland) con il metodo $Repeated-Weighting$ (RW). Questo metodo, che illustreremo brevemente nel paragrafo 2.1, può essere visto come un doppio passo di calibrazione (Traat e Särndal, 2011), in cui nel primo step si raggiunge la coerenza esterna con le P variabili ausiliarie X e nel secondo step con le M variabili Z.

Il metodo proposto (Ceccarelli *e altri*, 2010) prevede un unico passo di calibrazione, in cui si impone il vincolo di coerenza contemporaneamente rispetto a $\mathbf{t_X}$ e a $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ (cfr. anche Ceccarelli *e altri*, 2011; Traat e Särndal, 2011; Ceccarelli e Guandalini, 2013). I risultati più importanti sono relativi alla formulazione dell'espressione dello stimatore della varianza campionaria che, come sarà ampiamente spiegato, è di fondamentale importanza per valutare l'impiego di queste informazioni ausiliarie. Nel paragrafo 2.2.1 sarà presentata l'espressione valida per il caso di indagini indipendenti, mentre nel paragrafo 2.2.2 per il caso di indagini dipendenti.

2.1 Il *Repeated-weighting*

Il CBS ha messo a punto un metodo, *Repeated-Weighting* (RW), per costruire un sistema integrato di indagini. Il loro principale obiettivo è quello di garantire la consistenza numerica attraverso delle tavole ottenute combinando insieme registri - fonti amministrative o censuarie - ed indagini (Kroese e Ressen, 1999; van Duin e Snijder, 2010; Houbiers, 2004; Knottnerus e van Duin, 2006).

Il metodo RW è stato sviluppato nel caso in cui una o più variabili sono disponibili da più fonti. Si basa sull'applicazione reiterata dello stimatore $GREG$ (o equivalentemente dello stimatore CAL). Per ciascuna tabella costruita viene determinato un sistema di pesi che garantisce la coerenza con le marginali delle tabelle precedentemente calcolate. Questo metodo consente di migliorare la consistenza, in quanto consente di ottenere stime più precise di quelle che possono essere ottenute con lo stimatore $GREG$ (o CAL) dal momento che utilizza un gran numero di informazioni ausiliarie da altre indagini.

L'applicazione del metodo si basa su alcune assunzioni:

- le indagini ed i registri devono avere lo stesso periodo di riferimento;
- le indagini ed i registri devono riferirsi alla stessa popolazione;
- le variabili con lo stesso nome hanno la stessa definizione;
- le variabili categoriali hanno una classificazione gerarchica.

Se le prime tre condizioni non sono verificate non ha senso applicare il metodo *RW* in quanto si imporrebbe la coerenza tra quantità che non sono necessariamente coerenti tra loro.

La procedura di stima con il *RW* consiste in tre passi:

1. individuazione e ordinamento delle tavole che si intendono stimare;
2. stima delle tavole attraverso lo stimatore *GREG*;
3. aggiustamento iterativo delle tavole.

Per illustrare in maniera più chiara il metodo, riportiamo l'esempio fornito da van Duin e Snijder (2010) adottando la notazione già utilizzata nei precedenti paragrafi.

Ipotizziamo di avere due indagini basate su due campioni, s_1 e s_2, non sovrapposti. L'indagine su s_1 rileva la variabile X e la variabile Z, mentre quella svolta sul campione s_2 rileva anche la Y.

Da questo set di microdati possono essere creati tre blocchi (passo 1). Ciascun blocco è un sottogruppo della popolazione in cui sono raccolte le unità su cui sono rilevate le stesse variabili. I blocchi sono ordinati in maniera decrescente in base al numero di unità che raccolgono. Rappresentano, dunque, il più grande sottoinsieme di unità dal quale è possibile stimare la tabella, in modo da contenere la varianza delle stime (per una rappresentazione schematica si veda in Figura 2.1).

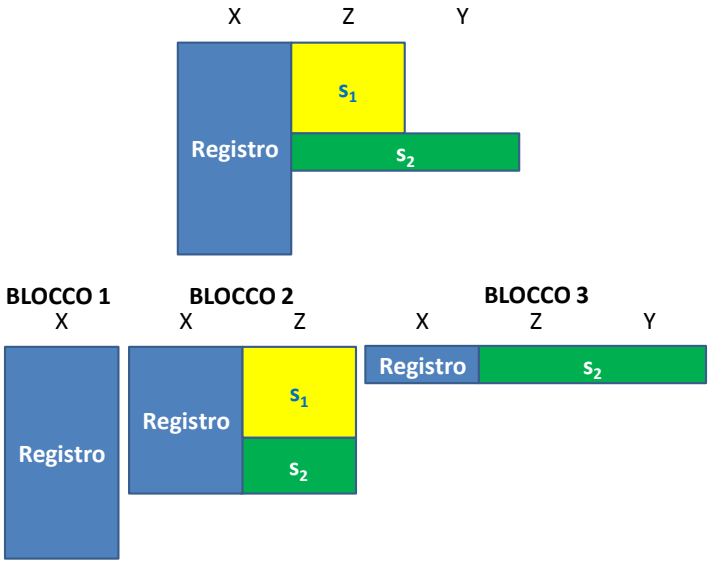
Il processo di stima delle tavole, passo 2, si articola in due passaggi. Il primo prevede la determinazione dei pesi di partenza (*starting weights*) per ciascun blocco individuato nel passo precedente. Nell'esempio, si ha che per il blocco 1 ogni unità avrà peso 1 in quanto composto solamente dal registro. Per il blocco 3 i pesi di partenza sono dati, invece, dai pesi utilizzati nell'indagine su s_2, in quanto composto solo da unità appartenenti a s_2. Nel blocco 2, invece, gli *starting weights*, d_k^2, sono determinati con l'espressione

$$d_k^{(2)} = \begin{cases} \frac{\lambda_1}{\pi_{k1}} & \text{se } k \in s_1 \text{ e } k \notin s_2 \\ \frac{(1-\lambda_1)}{\pi_{k2}} & \text{se } k \notin s_1 \text{ e } k \in s_2 \end{cases} \quad \text{con } \lambda_1 + \lambda_2 = 1$$

in modo da garantire che il sistema di pesi di partenza sia almeno coerente con il totale della popolazione. I valori λ_1 e λ_2 sono determinati tenendo conto della numerosità campionaria delle indagini oppure di un'approssimazione proposta da Kish (1992) (cfr. van Duin e Snijder, 2010, p. 8).

Una volta determinati i pesi di partenza si provvede al passaggio di *re-weighting* per le tavole che hanno delle marginali in comune con tavole basate su un maggior

Figura 2.1: Rappresentazione schematica del metodo *RW* quando s_1 e s_2 non sono sovrapposti (van Duin e Snijder, 2010, p. 6).



numero di unità e, quindi, stimate in precedenza (passo 3). Nell'esempio fornito questa necessità si presenta per il blocco 3, in quanto produce la tavola $[Z \times Y]$ in cui le marginali di Z sono state già calcolate dal blocco 2 che si basa su un numero di unità maggiore. Quindi, si procede a *ri-pesare* i pesi del blocco 3.

A questo punto possono essere seguite due strategie, una chiamata *minimal repeated weighting* e l'altra *splitting up procedure*. La prima prevede di calibrare i pesi del blocco in questione solamente sulle marginali che devono essere stimate. Sempre con riferimento all'esempio, i pesi del blocco 3 vengono calibrati sulle marginali delle Z stimate dal blocco 2, quindi, la coerenza con la X non è direttamente richiesta e, non essendo imposta, non viene necessariamente soddisfatta. Oltre a questo inconveniente la strategia *minimal repeated weighting* può portare a stime diverse a seconda della sequenza in cui vengono ordinate le tavole. In alcuni casi, infatti, l'ordine dei blocchi non è univocamente determinato e scelte differenti possono portare a stime leggermente differenti.

La *splitting up procedure* supera questi inconvenienti ricalibrando i pesi di ciascun blocco tenendo conto di tutte le marginali che possono essere prodotte dal blocco e che sono state precedentemente stimate. In pratica, per il blocco 3 viene soddisfatta la condizione di coerenza sia per la Z che per la X.

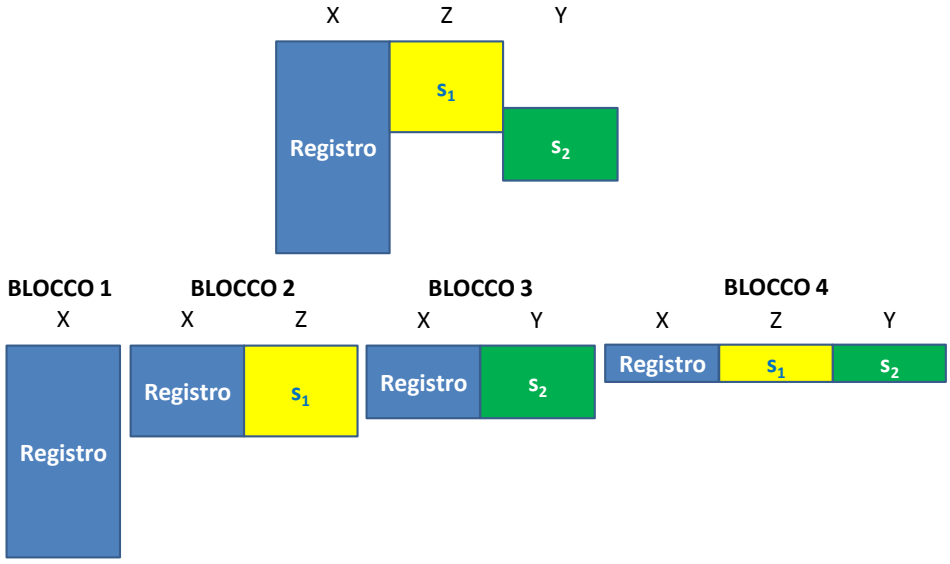
Per completezza, in Figura 2.2, si riporta il caso in cui i campioni s_1 e s_2 hanno delle unità in comune. Vengono, dunque, definiti 4 blocchi e i pesi di partenza per

il blocco 4, $d_k^{(4)}$, sono determinati con l'espressione

$$d_k^{(4)} = \begin{cases} \frac{\lambda_1}{\pi_{k1}} & \text{se } k \in s_1 \text{ e } k \notin s_2 \\ \frac{(1-\lambda_1)}{\pi_{k2}} & \text{se } k \notin s_1 \text{ e } k \in s_2 \quad \text{con } \lambda_1 + \lambda_2 = 1 \\ \frac{\lambda_1}{\pi_{k1}} + \frac{(1-\lambda_1)}{\pi_{k2}} & \text{se } k \in s_1 \bigcap s_2 \end{cases}$$

Nel passo di *re-weighting* si provvede a rendere i pesi del blocco 4 coerenti con i marginali della Z dal blocco 2 e della Y dal blocco 3.

Figura 2.2: Rappresentazione schematica del metodo RW quando s_1 e s_2 sono sovrapposti (van Duin e Snijder, 2010, p. 15).



Appare chiaro, dunque, che in situazioni più articolate il metodo RW comporta un aumento notevole della complessità del procedimento di stima in quanto richiede un numero maggiore di tavole da stimare e un numero maggiore di *ri-pesature* che possono essere spesso superflue. In alcuni casi la richiesta di coerenza con alcune marginali di pesi relativi a poche unità può portare a stime instabili e meno efficienti. Inoltre, per indagini su larga scala, il singolo processo di determinazione di pesi calibrati richiedere molto tempo-macchina. Il processo di *re-weighting*, che necessita di più processi di calibrazione, può richiedere, dunque, tempi particolarmente lunghi per ottenere delle stime.

L'espressione della varianza, illustrata da Knottnerus e van Duin (2006, p. 572), è funzione delle quantità che i due autori chiamano super-residui:

$$Var\left(\hat{t}_{Y_{RW}}\right) = \sum_{b=1}^{2} \sum_{k,l \in S_b} \left(\pi_{bkl} - \pi_{bk}\ \pi_{bl}\right) \frac{e_{bk}}{\pi_{bk}} \frac{e_{bl}^t}{\pi_{bl}},$$

in cui con S_b si indica l'insieme di unità che costituiscono il blocco b. Stimando i super-residui per s_1 e s_2

$$\begin{aligned}\hat{e}_{1k} &= \lambda_1 \hat{\beta}_{Y|Z} \left(z_k - x_k \hat{\beta}_{Z|X} \right) \\ \hat{e}_{2k} &= \left(y_k - z_k \hat{\beta}_{Y|Z} \right) - \lambda_1 \hat{\beta}_{Y|Z} \left(z_k - x_k \hat{\beta}_{Z|X} \right)\end{aligned}$$

e i coefficienti di regressione $\hat{\beta}_{Y|Z}$ e $\hat{\beta}_{Z|X}$ si può ottenere una stima della varianza attraverso l'espressione

$$var\left(\hat{t}_{Y_{RW}}\right) = \sum_{b=1}^{2} \sum_{k \in S_b} \left(d_k^{S_b} \right)^2 \ \hat{e}_{bk} \ \hat{e}_{bl}.$$

2.2 Lo stimatore calibrato con totali campionari

La metodologia *RW*, illustrata nel paragrafo precedente, ha come obiettivo primario quello di costruire un sistema di indagini coerenti tra loro.

Facciamo ora un passo indietro e consideriamo come obiettivo primario quello di includere nel processo di stima oltre al set di variabili ausiliarie X di cui si conoscono i totali da fonte amministrativa, anche un set di variabili ausiliarie Z per cui i totali di controllo sono stime. Assumiamo, quindi, di trovarci nello scenario ampiamente illustrato nel paragrafo 1.6.

Un modo per includere queste informazioni nel processo di stima è quello di individuare un sistema di pesi risolvendo il problema di minimo vincolato

$$\begin{cases} \min_{w_k} \left\{ \sum_{k \in s} G_k \left(w_k - d_k \right) / q_k \right\} \\ \sum_{k \in s} w_k \left(\mathbf{x}_k \ \mathbf{z}_k \right) = \left(\mathbf{t_X} \ \widetilde{\mathbf{t}}_{\mathbf{Z}_1} \right) \end{cases} \tag{2.1}$$

(Ceccarelli *e altri*, 2010).

A differenza del problema nella (1.22), nella (2.1) viene imposto il vincolo di coerenza oltre che sui totali $\mathbf{t_X}$ anche sui totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ (Ceccarelli *e altri*, 2010, 2011).

Il sistema di pesi finali applicato ai valori della variabili Y rilevati sul campione dà luogo allo stimatore

$$\hat{t}_{Y_{AC}} = \sum_{k \in s} y_k \ w_k = \mathbf{y}_s{}^t \ \mathbf{w}_s \tag{2.2}$$

che chiameremo stimatore AC adottando il nome attribuito da Traat e Särndal (2011). L'espressione (2.2) è esattamente equivalente a quella dello stimatore CAL.

Traat e Särndal (2011) affermano che nella metodologia *RW* il passo di *re-weighting* può essere visto come un ulteriore passo di calibrazione. Con lo stimatore

AC si ricorre ad un unico passo di calibrazione in cui la condizione coerenza dei pesi è imposta direttamente su entrambi i set di variabili X e Z.

Il vettore dei pesi finali che si ricava risolvendo il problema di minimo vincolato[1] nella (2.1) è

$$\mathbf{w}_{AC} = \mathbf{w}_{CAL} + \hat{\mathbf{R}}\ \mathbf{Z}_2 \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right)^{-1} \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{Z}_2{}^t\ \mathbf{w}_{CAL}\right), \tag{2.3}$$

in cui $\mathbf{w}_{CAL}$ è il vettore dei pesi finali definito nella (1.23) e ottenuti risolvendo il problema di minimo vincolato nella (1.22), mentre

$$\begin{aligned} \hat{\mathbf{R}} &= \mathbf{\Pi}_2^{-1} - \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2 \left(\mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\mathbf{X}_2\right) \mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1} \\ &= \mathbf{\Pi}_2^{-1} \left(\mathbf{I} - \hat{\mathbf{P}}\right) \end{aligned}$$

è la matrice dei residui campionari della regressione delle variabili Z_p, con $p = 1, \ldots, P$, sullo spazio generato dalle colonne della matrice $\mathbf{X}_2$. Si fa notare, infatti, che

$$\hat{\mathbf{P}} = \mathbf{X}_2 \left(\mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\mathbf{X}_2\right) \mathbf{X}_2{}^t \mathbf{\Pi}_2^{-1}$$

è la matrice degli operatori di proiezione sullo spazio generato dalle colonne della matrice $\mathbf{X}_2$.

La condizione di coerenza con i totali di controllo campionari, $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$, è ottenuta come residuale rispetto a quella dei totali di controllo, $\mathbf{t}_{\mathbf{X}}$.

In ogni caso i pesi finali nella (2.3) consentono di soddisfare la condizione di coerenza per entrambi i set di variabili. Infatti, se si stimano con *AC* i totali delle P variabili ausiliarie X sul campione s_2, si dimostra che

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{X}_{AC}} &= \mathbf{X}_2{}^t\ \mathbf{w}_{AC} \\ &= \mathbf{X}_2{}^t\ \mathbf{w}_{CAL} + \mathbf{X}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2 \left(\mathbf{X}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right)^{-1} \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{Z}_2{}^t\ \mathbf{w}_{CAL}\right) \\ &= \mathbf{t}_{\mathbf{X}} \end{aligned}$$

in quanto $\mathbf{X}_2{}^t\ \mathbf{R} = \underline{\mathbf{0}}$ e se si stimano i totali delle M variabili ausiliarie Z

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{Z}_{AC}} &= \mathbf{Z}_2{}^t\ \mathbf{w}_{AC} \\ &= \mathbf{Z}_2{}^t\ \mathbf{w}_{CAL} + \mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2 \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right)^{-1} \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{Z}_2{}^t\ \mathbf{w}_{CAL}\right) \\ &= \widetilde{\mathbf{t}}_{\mathbf{Z}_1}. \end{aligned}$$

Una prassi frequente, quando vengono usati dei totali di controllo campionari come in questo caso, è quella di impiegare le espressioni degli stimatori della varianza

[1] Si veda dimostrazione in Appendice, A3.

visti nel capitolo 1 sotto la tacita assunzione che l'errore associato ai totali di controllo campionari ha un impatto trascurabile sull'errore delle stime prodotte e può essere ignorato. Berger *e altri* (2009); Dever e Valliant (2010), però, dimostrano che con questi stimatori si sottostima notevolmente il reale valore della varianza dello stimatore.

2.2.1 La stima della varianza nel caso di indagini indipendenti

L'espressione dello stimatore della varianza per lo stimatore AC che proponiamo è determinata sfruttando ancora lo stretto legame che lega lo stimatore CAL e lo stimatore $GREG$. Lo stimatore AC, infatti, asintoticamente è equivalente allo stimatore AR, *Adjusted Regression* (Renssen e Nieuwenbroek, 1997), in letteratura chiamato anche stimatore *Extended regression* (E) (Ballin *e altri*, 2000; Rancourt, 2001; Berger *e altri*, 2009).

Lo stimatore AR può essere scritto come

$$\hat{t}_{Y_{AC}} \asymp \hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\check{\mathbf{t}}_{\mathbf{U}} - \hat{\mathbf{t}}_{\mathbf{U}_{HT}}\right)^t \hat{\boldsymbol{\beta}} \tag{2.4}$$

(Ceccarelli e Guandalini, 2013) dove

- $\check{\mathbf{t}}_{\mathbf{U}}{}^t = \left(\mathbf{t}_{\mathbf{X}}{}^t \; \widetilde{\mathbf{t}}_{\mathbf{Z}_1}{}^t\right)$ è il vettore dei totali di controllo, in cui i primi P elementi sono i totali di controllo non-campionari e gli ultimi M sono totali di controllo campionari stimati su s_1;
- $\hat{\boldsymbol{\beta}}^t = \left(\hat{\boldsymbol{\beta}}_P{}^t \; \hat{\boldsymbol{\beta}}_M{}^t\right)$ è il vettore delle stime dei coefficienti di regressione, $\boldsymbol{\beta}$, che contiene $M + P$ elementi, dove $\hat{\boldsymbol{\beta}}_P$ sono i primi P coefficienti di regressione relativi al set di variabili X e $\hat{\boldsymbol{\beta}}_M$ gli ultimi M relativi al set di variabili Z;
- $\check{\mathbf{t}}_{\mathbf{U}_{HT}}{}^t = \left(\mathbf{t}_{\mathbf{X}_{HT}}{}^t \; \widetilde{\mathbf{t}}_{\mathbf{Z}_{HT}}{}^t\right)$ è il vettore dei totali stimato su s_2 con lo stimatore HT composto da $M + P$ elementi, $q = 1, \ldots, M + P$.

Aggiungendo e sottraendo la quantità $\mathbf{t}_{\mathbf{U}}{}^t\boldsymbol{\beta}$ nella (2.4), dove

$$\mathbf{t}_{\mathbf{U}}{}^t = \left(\mathbf{t}_{\mathbf{X}}{}^t \; \mathbf{t}_{\mathbf{Z}}{}^t\right)$$

contiene i totali della popolazione anche per il set di variabili Z, la (2.4) può essere scritta in maniera equivalente come

$$\begin{aligned} \hat{t}_{Y_{AR}} &= \hat{t}_{Y_{HT}} + \left(\check{\mathbf{t}}_{\mathbf{U}} - \mathbf{t}_{\mathbf{U}} + \mathbf{t}_{\mathbf{U}} - \hat{\mathbf{t}}_{\mathbf{U}_{HT}}\right)^t \hat{\boldsymbol{\beta}} \\ &= \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{U}} - \hat{\mathbf{t}}_{\mathbf{U}_{HT}}\right)^t \hat{\boldsymbol{\beta}} - \left(\mathbf{t}_{\mathbf{U}} - \check{\mathbf{t}}_{\mathbf{U}}\right)^t \hat{\boldsymbol{\beta}}. \end{aligned}$$

Quest'espressione può essere scissa in due quantità

$$A_1 = \hat{t}_{Y_{HT}} + \left(\mathbf{t_U} - \hat{\mathbf{t}}_{\mathbf{U}_{HT}}\right)^t \hat{\boldsymbol{\beta}},$$

che coincide con lo stimatore *GREG* definito nella (1.11) nel caso in cui si hanno due set di informazioni ausiliarie X e Z con totali noti senza errore campionario, mentre

$$A_2 = \left(\mathbf{t_U} - \check{\mathbf{t}}_{\mathbf{U}}\right)^t \hat{\boldsymbol{\beta}}$$

è la *distorsione* che si commette nella stima considerando totali di controllo stimati da s_1, $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$, invece che ricorrendo al valore vero nella popolazione, $\mathbf{t_Z}$.

La varianza dello stimatore in (2.4), quindi, può essere scritta come

$$Var\left(\hat{t}_{Y_{AR}}\right) = Var(A_1) + Var(A_2) - 2\, Cov(A_1, A_2) \tag{2.5}$$

La varianza di A_1, ovvero la varianza dello stimatore *GREG*, è data dalla (1.13) e può essere stimata attraverso lo stimatore già illustrato nella (1.14). La varianza di A_2 è

$$Var\left(A_2\right) = \left({\boldsymbol{\beta}_P}^t \; {\boldsymbol{\beta}_M}^t\right) \begin{bmatrix} Var\left(\mathbf{t_X}\right) & Cov\left(\mathbf{t_X}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \\ Cov\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}, \mathbf{t_X}\right) & Var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M \end{pmatrix}$$

e un suo stimatore è dato da

$$var\left(A_2\right) = \left({\hat{\boldsymbol{\beta}}_P}^t \; {\hat{\boldsymbol{\beta}}_M}^t\right) \begin{bmatrix} 0 & 0 \\ 0 & var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}_P \\ \hat{\boldsymbol{\beta}}_M \end{pmatrix}$$

Questa espressione, poiché la matrice di varianze e covarianze campionarie di $\check{\mathbf{t}}_{\mathbf{U}}$ ha una struttura a blocchi, in quanto $\mathbf{t_X}$ è una costante rispetto al campione, si riduce a

$$\begin{aligned} var\left(A_2\right) &= {\hat{\boldsymbol{\beta}}_M}^t \; var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \; \hat{\boldsymbol{\beta}}_M \\ &= \sum_{m=1}^{M} {\hat{\beta}_m}^2 \; var\left(\hat{t}_{Z_{1m}}\right) + \sum_{m' \neq m} \hat{\beta}_{m'} \hat{\beta}_m \; cov\left(\hat{t}_{Z_{1m'}}, \hat{t}_{Z_{1m}}\right) \end{aligned}$$

(cfr. Ceccarelli *e altri*, 2010, 2013).
Nella varianza di A_2 compaiono i coefficienti di regressione stimati su s_2 relativi a ciascuna variabile Z_m, con $m = 1, \ldots, M$, e le stime delle varianze e covarianze campionarie del set di variabili Z.

Un discorso più complesso, invece, riguarda la $Cov(A_1, A_2)$. Nel caso in cui i campioni s_1 e s_2 sono estratti in maniera indipendente non vi è sovrapposizione tra i due campioni[2] e $Cov(A_1, A_2) = 0$.

Uno stimatore della (2.5) nel caso di indagini indipendenti è dato, dunque, da

$$\begin{aligned} var\left(\hat{t}_{Y_{AC}^{ind}}\right) &= var\left(\hat{t}_{Y_{GREG}}\right) + \hat{\boldsymbol{\beta}}_M{}^t \; var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \; \hat{\boldsymbol{\beta}}_M \\ &= \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_k \; \hat{e}_k\right) \left(\gamma_l \; \hat{e}_l\right) \\ &+ \sum_{m=1}^{M} \hat{\beta}_m{}^2 \; var\left(\hat{t}_{Z_{1m}}\right) + \sum_{m' \neq m} \hat{\beta}_{m'} \; \hat{\beta}_m \; cov\left(\hat{t}_{Z_{1m'}}, \hat{t}_{Z_{1m}}\right) \end{aligned} \tag{2.6}$$

(cfr. Ceccarelli *e altri*, 2010, 2013).
Nella (2.6) il primo elemento, $var\left(\hat{t}_{Y_{GREG}}\right)$, rappresenta l'errore che si commetterebbe qualora il vettore dei totali di controllo relativi alle M variabili Z non fosse affetto da errore campionario, cioè se si conoscesse $\mathbf{t_Z}$. Il secondo elemento, invece, tiene conto del fatto che in realtà è un vettore di stime, $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$, e quindi affetto da errore campionario.

La seconda parte della (2.6) è una quantità sempre positiva, in quanto è una forma quadratica semi-definita positiva e ci consente di quantificare l'errore che si importa nelle stime quando si ricorre a totali di controllo campionari. Questo risultato, dunque, è in completo accordo con quanto sostenuto da Berger *e altri* (2009); Dever e Valliant (2010), in quanto dimostra che considerando solo il primo termine della (2.6) si sottostimerebbe la varianza dello stimatore $\hat{t}_{Y_{AC}^{ind}}$.

2.2.2 La stima della varianza nel caso di indagini dipendenti

Quando i campioni s_1 e s_2 non sono estratti in maniera indipendente, $Cov(A_1, A_2) \neq 0$ e deve essere stimata. L'impiego di totali noti campionari nel caso di campioni dipendenti è molto diffuso ed è particolarmente frequente nel caso di indagini con campioni ruotati il cui obiettivo è di produrre stime sia trasversali che longitudinali, oltre che di stimare le differenze tra due periodi.

Nel caso di disegni campionari ruotati, ricorrere ad informazioni ausiliarie stimate in occasioni di rilevazione precedenti (informazioni ausiliarie longitudinali) consente

[2]Nel caso in cui i campioni s_1 e s_2 sono estratti indipendentemente la probabilità che abbiano delle unità in comune è prossima allo 0, ma non uguale a 0. Può capitare, quindi, che vi sia una piccolissima sovrapposizione tra i due campioni, ma questa può essere trascurata. Ovviamente questa valutazione è legata alla dimensione della popolazione, N, e alle dimensioni di s_1 ed s_2 rispetto alla popolazione.

di ridurre la variabilità delle stime sia di livello che di cambiamento riducendo notevolmente la volatilità delle serie stimate (Singh *e altri*, 2001). Nel caso di campioni ruotati è frequente che i totali di controllo desunti da s_1 siano le stime ottenute nell'occasione precedente di indagine dei parametri che si vogliono stimare ora su s_2. Poiché si basano su una frazione $q(s_C)$, $0 \leq q(s_C) \leq 1$, di unità rilevate in entrambe le rilevazioni, le stime saranno fortemente correlate tra loro e proprio questo consente di ridurre l'errore campionario.

Nella letteratura gli stimatori che considerano informazioni ausiliarie longitudinali sono definiti *composite estimator* (Wolter, 1979). Diversi sviluppi relativi a questa classe di stimatori sono stati ottenuti nell'ambito della rilevazione mensile sulle Forze di Lavoro canadese (Canadian Labour Force Survey, CLFS). Ad esempio sono stati definiti lo stimatore composito K, lo stimatore composito AK, lo stimatore $MR2$ e lo stimatore AK^* (cfr. Fuller e Rao, 2001; Singh *e altri*, 2001, per una panoramica). In particolare lo stimatore AK^*, impiegato dal 2000 nella CLFS, è una combinazione lineare di stimatori di tipo $GREG$ che si basano sulla frazione di unità sovrapposte, $q(s_C)$, e sulla media delle unità non sovrapposte chiamate *Birth* (B) se sono osservate per la prima volta al tempo t o *Death* (D) se, dopo essere state osservate per l'ultima volta al tempo $t-1$, sono uscite dallo schema di rotazione.

Questo tipo di stimatori è particolarmente adatto al disegno della CLFS che prevede uno schema di rotazione con un'elevata sovrapposizione tra i campioni s_1 e s_2 ($q(s_C) = 5/6$). È, tuttavia, meno adatto al caso in cui la sovrapposizione è minore e qualora si vogliano inserire totali longitudinali riferiti a più periodi precedenti, come nel caso della Rilevazione Continua delle Forze Lavoro italiana (RCFL). Inserire queste informazioni ausiliarie longitudinali nel problema di minimo vincolato visto nella (2.1) consente di utilizzarle nel processo di stima in maniera più semplice ed elastica.

Lo sviluppo dello stimatore segue esattamente quanto detto sinora. Come già accennato, il problema metodologico in questo caso risiede nella determinazione dell'espressione dello stimatore della varianza e in particolar modo nella componente di $Cov(A_1, A_2)$ nella (2.5).

Assumiamo che le stime di $\mathbf{t_Z}$ ottenute su s_1 siano non-distorte. In formule

$$\mathbb{E}\left[\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right] = \mathbf{t_Z}.$$

Questa ipotesi è realistica perché generalmente nelle indagini vengono utilizzati stimatori che godono della proprietà di non-distorsione. Di conseguenza

$$\begin{aligned} \mathbb{E}\left[A_2\right] &= \mathbb{E}\left[\left(\mathbf{t_U} - \widetilde{\mathbf{t}}_{\mathbf{U}}\right)^t \boldsymbol{\beta}\right] \\ &= \mathbf{0}. \end{aligned}$$

Nell'espressione precedente si è considerato $\boldsymbol{\beta}$ e non $\hat{\boldsymbol{\beta}}$, in quanto si è fatto ricorso ad una semplificazione spesso utilizzata nel caso in cui si opera con lo stimatore $GREG$, ovvero che i coefficienti di regressione siano delle costanti. Questa semplificazione si basa sulla dimostrazione che la differenza tra $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{\beta}$ è una quantità di ordine $n^{-1/2}$ e diventa trascurabile al crescere della dimensione del campione (cfr, e.g., Conti e Marella, 2011, p. 88).

Questa semplice assunzione consente di scrivere[3]

$$\begin{aligned} Cov(A_1, A_2) &= \mathbb{E}\left[A_1\ A_2\right] \\ &= \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \left(\mathbf{t_U} - \check{\mathbf{t}}_{\mathbf{U}}\right)^t \boldsymbol{\beta}\right] \\ &= -Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M \\ &= \boldsymbol{\beta}^t\ Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M - Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M \qquad (2.7) \end{aligned}$$

in quanto $\mathbb{E}\left[\hat{t}_{Y_{GREG}}\right] = \hat{t}_Y$ almeno asintoticamente e $\mathbf{t_X}$ è una costante rispetto al campione.

Con riferimento alla (2.7), $Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ è una matrice di $(M + P) \times P$ covarianze, mentre $Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ è un vettore di P covarianze.

Gli elementi che compongono la matrice $Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ ed il vettore $Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ sono covarianze tra due stimatori di totali ottenuti da due indagini differenti che però hanno in comune una quota $q(s_C)$ di unità. Una soluzione per giungere ad una stima di queste quantità, anche nel caso di disegni campionari complessi, è stata ottenuta estendendo alcuni risultati forniti dal lavoro di Qualité e Tillé (2008).

I due autori propongono un'espressione della stima della varianza delle differenze in campioni ruotati nel caso di campionamento semplice e stratificato in cui, come nel nostro caso, è richiesta la stima della covarianza tra due stimatori di totali ottenuti da due indagini differenti che però hanno in comune una quota di unità. Lo stimatore che propongono è in grado di tener conto di alcune componenti di questi stimatori tra cui il disegno campionario, il trattamento per mancata risposta e la calibrazione. Illustriamo brevemente l'idea alla base dello stimatore delle covarianze proposto da Qualité e Tillé (2008).

I campioni s_1 e s_2 possono essere suddivisi in tre parti (cfr. anche Figura 2.3)

- $s_A = s_1 \setminus s_2$;
- $s_B = s_2 \setminus s_1$;

[3]Si veda dimostrazione in Appendice, A4.

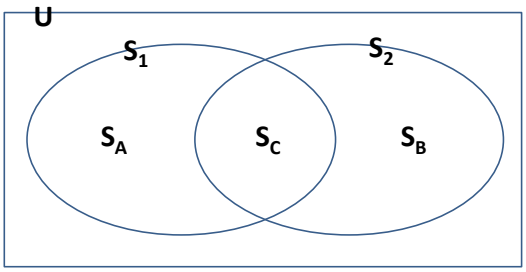
- $s_C = s_1 \cap s_2$.

Siano:

- $n_A = |s_A|$;
- $n_B = |s_B|$;
- $n_C = |s_C|$;
- $n_1 = n_A + n_C$;
- $n_2 = n_B + n_C$

dove s_A, s_B e s_C possono essere random.

Figura 2.3: Campioni sovrapposti (Qualité e Tillé, 2008, p. 174).



La distribuzione di probabilità di due campioni casuali s_1 e s_2 condizionatamente a n_A, n_B e n_C è data da:

$$p\left(s_1, s_2 | n_A, n_B, n_C\right) = \begin{cases} \frac{n_A!\ n_B!\ n_C!(N-n_A-n_B-n_C)!}{N!} & \begin{array}{r} n_A = |s_A| \\ se\ n_B = |s_B| \\ n_C = |s_C| \end{array} \\ 0 & altrimenti \end{cases},$$

(cfr. Goga, 2008).

Aggiungendo l'ipotesi che condizionatamente a n_A, n_B e n_C i campioni s_A, s_B e s_C sono casuali semplici senza ripetizione di dimensione fissata, la legge di probabilità per estrarre s_1 e s_2 non è conosciuta in generale, ma si può assumere essere della forma:

$$p(s_1, s_2) = p(s_1, s_2 | n_A, n_B, n_C)\ Pr(|s_1 \cap s_2| = n_C).$$

Indichiamo con X la variabile osservata solo sulle unità del campione s_1 e con Y quella osservata solo sulle unità del campione s_2 e notando che:

$$\bar{x}_A = \frac{1}{n_A}\sum_{k\in s_A} x_k, \qquad \bar{x}_C = \frac{1}{n_C}\sum_{k\in s_C} x_k,$$

$$\bar{y}_B = \frac{1}{n_B}\sum_{k\in s_B} y_k, \qquad \bar{y}_C = \frac{1}{n_C}\sum_{k\in s_C} y_k,$$

$$\bar{x}_1 = \frac{n_A\ \bar{x}_A + n_C\ \bar{x}_C}{n_1}, \qquad \bar{y}_2 = \frac{n_B\ \bar{y}_B + n_C\ \bar{y}_C}{n_1},$$

quindi $\hat{t}_X = N\bar{x}$ e $\hat{t}_Y = N\bar{y}$. La covarianza di $\bar{x}$ e $\bar{y}$ condizionatamente a n_A, n_B e n_C è data da:

$$\begin{aligned} cov\left(\bar{x},\bar{y}\right) &= \mathbb{E}\left[cov(\bar{x}_1,\bar{y}_2|n_A,n_B,n_C)\right] \\ &+ cov\left(\mathbb{E}\left[\bar{x}_1|n_A,n_B,n_C\right]\mathbb{E}\left[\bar{y}_2|n_A,n_B,n_C\right]\right) \end{aligned}$$

Poichè $\bar{x}_1$ e $\bar{y}_2$ sono non-distorte condizionatamente a n_A, n_B e n_C:

$$Cov\left(\mathbb{E}\left[\bar{x}_1|n_A,n_B,n_C\right]\mathbb{E}\left[\bar{y}_2|n_A,n_B,n_C\right]\right) = Cov\left(\bar{X},\bar{Y}\right) = 0.$$

e

$$Cov\left(\bar{x}_1,\bar{y}_2\right) = \mathbb{E}\left[cov(\bar{x}_1,\bar{y}_2|n_A,n_B,n_C)\right].$$

La covarianza condizionata (Tam, 1984, p. 289) è uguale a

$$Cov(\bar{x}_1,\bar{y}_2|n_A,n_B,n_C) = \left(\frac{n_c}{n_1\ n_2} - \frac{1}{N}\right) S_{xy}$$

e quindi

$$Cov(\bar{x}_1,\bar{y}_2) = \left(\frac{\mathbb{E}[n_C]}{n_1\ n_2} - \frac{1}{N}\right) S_{xy}$$

in cui $\mathbb{E}[n_C]$ è il valore atteso dal numero di unità in comune tra i campioni s_1 e s_2 e può essere sostituito con la quantità n_C, ovvero il numero osservato di unità in comune tra i due campioni. La covarianza di due stime nel caso di campionamento casuale semplice è, dunque,

$$cov(\hat{t}_X,\hat{t}_Y) = N^2\left(\frac{\mathbb{E}[n_C]}{n_1\ n_2} - \frac{1}{N}\right) S_{xy_C}$$

dove

$$S_{xy_C} = \frac{1}{N-1}\sum_{k\in U}\left(x_k - \bar{X}\right)\left(y_k - \bar{Y}\right)$$

può essere stimata sulle unità presenti in entrambe le occasioni di indagine da

$$\hat{s}_{xy_C} = \frac{1}{n_C-1}\sum_{k\in s_C}\left(x_k - \bar{x}_C\right)\left(y_k - \bar{y}_C\right)$$

e quindi una stima della covarianza campionaria è data da

$$cov(\hat{t}_X, \hat{t}_Y) = N^2 \left(\frac{\mathbb{E}[n_C]}{n_1 \; n_2} - \frac{1}{N} \right) \hat{s}_{xy}. \tag{2.8}$$

Qualité e Tillé (2008) propongono anche un'espressione della covarianza valida per il caso del disegno stratificato. La (2.8) è, infatti, facilmente estendibile al caso in cui la popolazione è suddivisa in L strati ($h = 1, ..., L$) di numerosità N_h ($\sum_{h=1}^{L} N_h = N$) da cui si selezionano con disegno casuale semplice, in maniera indipendente all'interno di ciascuno strato, n_h unità ($\sum_{h=1}^{L} n_h = n$). L'unica assunzione necessaria è che le unità non cambino strato tra una rilevazione e l'altra. Indicando con n_{h_1} e n_{h_2} le unità estratte nello strato h-mo rispettivamente nella prima e nella seconda rilevazione, la covarianza tra $\hat{t}_X$ e $\hat{t}_Y$ all'interno del generico strato h è data da:

$$Cov_h(\hat{t}_X, \hat{t}_Y) = N_h^2 \left(\frac{\mathbb{E}[n_{h_C}]}{n_{h_1} \; n_{h_2}} - \frac{1}{N_h} \right) S_{xy_{h_C}}.$$

Inserendo la stima di S_{xy_C},

$$\hat{s}_{xy_{h_C}} = \frac{1}{n_{h_C} - 1} \sum_{k \in s_{h_c}} \left(x_{hk} - \bar{x}_{h_C} \right) \left(y_{hk} - \bar{y}_{h_C} \right),$$

nell'espressione precedente si ha la stima della covarianza, $cov_h(\hat{t}_X, \hat{t}_Y)$ nello strato h-mo. La stima nella popolazione sarà data, invece, da:

$$cov(\hat{t}_X, \hat{t}_Y) = \sum_{h=1}^{L} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \; cov_h(\hat{t}_X, \hat{t}_Y).$$

Facendo alcune semplici assunzioni abbiamo esteso il metodo proposto da Qualité e Tillé (2008) nel caso di disegno campionario a grappoli (o a uno stadio) e a due stadi. In questo modo possiamo stimare le covarianze e determinare la stima della varianza campionaria di AC nel caso di indagini dipendenti anche per disegni campionari complessi.

Nel caso del disegno ad uno stadio, l'idea è quella di riportare le metodologia vista per il disegno casuale semplice senza ripetizione delle unità all'estrazione dei grappoli.

Nel disegno a grappoli, infatti, la popolazione è suddivisa in M grappoli mutuamente esclusivi ($G = 1, ..., M$). Si estraggono con disegno casuale semplice, ipotizziamo senza ripetizione, m grappoli ($g = 1, ..., M$) e all'interno di ciascun grappolo selezionato si rilevano le informazioni su tutte le unità appartenenti al grappolo stesso.

Se sono verificate queste due assunzioni:

- la composizione dei grappoli resta invariata tra una rilevazione e l'altra;

- il numero di grappoli in comune tra le due rilevazione deve essere maggiore o uguale a 2, $m_C = m_1 \cap m_2 \geq 2$ ($g_c = 1, ..., m_C$), dove m_1 e m_2 sono i grappoli estratti rispettivamente nella prima e nella seconda rilevazione,

$$Cov(\hat{t}_X, \hat{t}_Y) = M^2 \left(\frac{m_C}{m_1\ m_2} - \frac{1}{M} \right) S_{xy_{b_C}},$$

dove $S_{xy_{b_C}}$ è la variabilità congiunta della t_Y e della t_X tra grappoli che può essere stimata con:

$$\hat{s}_{xy_{b_C}} = \frac{1}{m_C - 1} \sum_{g \in g_C} \left(\hat{t}_{X_g} - \frac{1}{m_C} \sum_{g \in g_C} \hat{t}_{X_g} \right) \left(\hat{t}_{Y_g} - \frac{1}{m_C} \sum_{g \in g_C} \hat{t}_{Y_g} \right).$$

Sostituendo, ovviamente, questa espressione in quella precedente si ottiene una stima della covarianza campionaria.

Nel caso di disegno campionario a due stadi con stratificazione delle unità di primo stadio si hanno L strati ($h = 1, ..., L$). In ciascuno strato, dove vi sono N_h unità di primo stadio, UPS ($i = 1, ..., N_h$), ne vengono estratte con disegno casuale semplice senza ripetizione n_h ($i = 1, ..., n_h$) con probabilità proporzionale alla ampiezza. La generica PSU contiene M_{hi} unità di secondo stadio, USS ($j = 1, ..., M_{hi}$), da cui vengono selezionate con disegno casuale semplice senza ripetizione m_{hi} USS ($j = 1, ..., m_{hi}$). Si ha, inoltre, $M_h = \sum_{i=1}^{n_h} M_{hi}$.

Quando è prevista una rotazione delle unità tra le rilevazioni, le unità in comune tra una rilevazione e l'altra si trovano sempre a livello del secondo stadio, mentre le unità di primo stadio sono comunque in comune tra le due rilevazioni. La maggior parte della variabilità, dunque, è ascrivibile alle USS (cfr. anche Singh *e altri*, 2001, p. 37). Avendo

- $N_{h_1} = N_{h_2} = N_h$;

- $n_{h_1} = n_{h_2} = n_h$;

- $M_{hi_1} = M_{hi_2} = M_{hi}$;

- m_{hij_1} sono le unità rilevate nella j-ma USS nella i-ma UPS nello strato h-mo nella prima rilevazione;

- m_{hij_2} sono le unità rilevate nella j-ma USS nella i-ma UPS nello strato h-mo nella seconda rilevazione;

- m_{hij_C} sono le unità in comune tra la prima e la seconda rilevazione nella j-ma USS nella i-ma UPS nello strato h-mo.

è possibile, dunque, estendere l'espressione della covarianza anche al caso del campionamento a due stadi.

La stima della varianza in un disegno a due stadi con stratificazione delle UPS può essere scritta come

$$var(\hat{t}_X, \hat{t}_Y) = var_1(\hat{t}_X, \hat{t}_Y) + var_2(\hat{t}_X, \hat{t}_Y),$$

ovvero la somma della stima della varianza campionaria al primo stadio:

$$var_1(\hat{t}_X, \hat{t}_Y) = N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \hat{s}_{T_{xy}},$$

dove

$$\hat{s}_{T_{xy}} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{t}_{X_{hi}} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{t}_{X_{hi}}\right) \left(\hat{t}_{Y_{hi}} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{t}_{Y_{hi}}\right)$$

e della stima della varianza campionaria al secondo stadio

$$var_2(\hat{t}_X, \hat{t}_Y) = \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right) \hat{s}_{2_{xy}},$$

dove

$$\hat{s}_{2_{xy}} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} \left(x_{hij} - \frac{1}{m_{hij}} \sum_{j=1}^{mhij} x_{hij}\right) \left(y_{hij} - \frac{1}{m_{hij}} \sum_{j=1}^{mhij} y_{hij}\right).$$

Il calcolo della covarianza, per quanto detto in precedenza, riguarderebbe le unità che, facenti parti della stessa UPS e della stessa USS, vengono rilevate in entrambe le occasioni. È possibile, quindi, scrivere la stima della covarianza in un disegno a due stadi con stratificazione delle UPS, a partire dall'espressione della stima della varianza come:

$$cov_h(\hat{t}_X, \hat{t}_Y) = cov_{h_1}(\hat{t}_X, \hat{t}_Y) + cov_{h_2}(\hat{t}_X, \hat{t}_Y).$$

Definendo:

- $\hat{t}_{X_{hi_C}} = \sum_{j=1}^{m_{hi_C}} \frac{M_{hi}}{m_{hi_C}} x_{hij_C}$;
- $\bar{x}_{h_C} = \frac{1}{N_h} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi_C}} \frac{M_h}{M_{hi}} \frac{M_{hi}}{m_{hi_C}} \frac{1}{n_h} x_{hij_C}$;
- $\hat{t}_{Y_{hi_C}} = \sum_{j=1}^{m_{hi_C}} \frac{M_h i}{m_{hi_C}} y_{hij_C}$;
- $\bar{y}_{h_C} = \frac{1}{N_h} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi_C}} \frac{M_h}{M_{hi}} \frac{M_{hi}}{m_{hi_C}} \frac{1}{n_h} y_{hij_C}$;
- $\bar{x}_{hi_C} = \frac{1}{m_{hi_C}} \sum_{j=1}^{m_{hi}} x_{hij_C}$;

- $\bar{y}_{hi_C} = \frac{1}{m_{hi_C}} \sum_{j=1}^{m_{hi}} y_{hij_C}$

e considerando che le USS vengono estratte con un disegno casuale semplice senza ripetizione, si può scrivere la stima della covarianza campionaria al primo stadio come

$$cov_{h_1}(\hat{t}_X, \hat{t}_Y) = N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \hat{s}_{T_{xy_C}},$$

dove

$$\hat{s}_{T_{xy_C}} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{t}_{X_{hi_C}} - \bar{x}_{h_C} \right) \left(\hat{t}_{Y_{hi_C}} - \bar{Y}_{h_C} \right)$$

e la stima della covarianza campionaria al secondo stadio come

$$cov_{h_2}(\hat{t}_X, \hat{t}_Y) = \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{\mathbb{E}_{hi_c}]}{m_{hi_1}\ m_{hi_2}} - \frac{1}{M_{hi}} \right) \hat{s}_{2_{xy}},$$

dove

$$\hat{s}_{2_{xy_C}} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} \left(x_{hij} - \bar{x}_{hi_C} \right) \left(y_{hij} - \bar{y}_{hi_C} \right).$$

La stima della covarianza nella popolazione sarà, quindi, data da:

$$cov(\hat{t}_X, \hat{t}_Y) = \sum_{h=1}^{L} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1}\ cov_h(\hat{t}_X, \hat{t}_Y).$$

Come accennato in precedenza Qualité e Tillé (2008, p. 177) forniscono anche un'approssimazione nel caso in cui si vuole studiare la correlazione tra stimatori di tipo CAL, quindi esattamente il nostro caso.

In particolare, poiché la varianza degli stimatori di tipo CAL è funzione dei residui e può essere approssimata calcolando la varianza dello stimatore HT del totale dei residui, propongono di sostituire al posto del valore, x_k o y_k registrato sulla generica unità campionaria, l'espressione dei residui

$$e_{k_1} = x_k - \mathbf{z}'_{k_1} \hat{\boldsymbol{\beta}}_1$$

$$e_{k_2} = y_k - \mathbf{z}'_{k_2} \hat{\boldsymbol{\beta}}_2$$

dove $\mathbf{z}_{k_t}$, con $t = 1, 2$, è un vettore colonna di variabili dummy utilizzate come variabili di controllo rispettivamente dagli stimatori $\hat{t}_X$ e $\hat{t}_Y$ non necessariamente uguali e

$$\hat{\boldsymbol{\beta}}_1 = \left(\sum_{k \in s_1} \frac{q_{k_1} \mathbf{z}_{k_1} \mathbf{z}'_{k_1}}{\pi_{k_1}} \right)^{-1} \left(\sum_{k \in s_1} \frac{q_{k_1} \mathbf{z}_{k_1} x'_{k_1}}{\pi_{k_1}} \right),$$

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{k \in s_2} \frac{q_{k_2} \mathbf{z}_{k_2} \mathbf{z}'_{k_2}}{\pi_{k_2}} \right)^{-1} \left(\sum_{k \in s_2} \frac{q_{k_2} \mathbf{z}_{k_2} x'_{k_2}}{\pi_{k_2}} \right),$$

con π_{tk} probabilità di inclusione dell'unità k nell'occasione di rilevazione t.

Nella (2.7) questa approssimazione è necessaria solamente per i totali di controllo campionari, $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ e ci consente di stimare direttamente gli elementi del vettore $Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ e della matrice $Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$.

Ad esempio, nel caso in cui s_1 e s_2 sono estratti con disegno casuale semplice senza ripetizione, le stime delle covarianze per ciascuno dei generici elementi, per $p = 1, \ldots, P$ e $m = 1, \ldots, M$, sarebbero

$$cov(\hat{t}_{Y_{HT}}, \widetilde{t}_{Z_1 m}) = N^2 \left(\frac{\mathbb{E}[n_C]}{n_1\ n_2} - \frac{1}{N} \right) \hat{s}_{YZ_C} \tag{2.9}$$

con

$$\hat{s}_{YZ_C} = \frac{1}{n_c - 1} \sum_{k \in s_C} \left(y_k - \frac{1}{n_2} \sum_{k \in s_C} y_k \right) \left(e_{k_1} - \frac{1}{n_1} \sum_{k \in s_C} e_{k_1} \right).$$

$$cov(\hat{t}_{X_{p_{HT}}}, \widetilde{t}_{Z_1 m}) = N^2 \left(\frac{\mathbb{E}[n_C]}{n_1\ n_2} - \frac{1}{N} \right) \hat{s}_{XZ_C} \tag{2.10}$$

con

$$\hat{s}_{XZ_C} = \frac{1}{n_c - 1} \sum_{k \in s_C} \left(x_k - \frac{1}{n_2} \sum_{k \in s_C} x_k \right) \left(e_{k_1} - \frac{1}{n_1} \sum_{k \in s_C} e_{k_1} \right).$$

e

$$cov(\hat{t}_{Z_{m_{HT}}}, \widetilde{t}_{Z_1 m}) = N^2 \left(\frac{\mathbb{E}[n_C]}{n_1 n_2} - \frac{1}{N} \right) \hat{s}_{ZZ_C} \tag{2.11}$$

con

$$\hat{s}_{ZZ_C} = \frac{1}{n_c - 1} \sum_{k \in s_C} \left(z_k - \frac{1}{n_2} \sum_{k \in s_C} z_k \right) \left(e_{k_1} - \frac{1}{n_1} \sum_{k \in s_C} e_{k_1} \right).$$

Dalla (2.9), (2.10) e (2.11) le espressioni nel caso di disegno stratificato, a grappolo e a due stadi possono essere ricavate in maniera semplice sulla base delle espressioni presentate in precedenza.

In base al disegno con cui sono estratti i campioni s_1 e s_2 siamo, dunque, in grado di stimare gli elementi che compongono il vettore $Cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ e la matrice $Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right)$ e sostituendoli nella (2.7), abbiamo che

$$\begin{aligned} cov(A_1, A_2) &= \mathbb{E}\left[A_1\ A_2\right] \\ &= \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \left(\mathbf{t}_{\mathbf{U}} - \check{\mathbf{t}}_{\mathbf{U}}\right)^t \boldsymbol{\beta}\right] \\ &= \boldsymbol{\beta}^t\ cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M - cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M. \end{aligned} \tag{2.12}$$

Dunque, uno stimatore della varianza campionaria di AC nel caso di indagini dipendenti è dato da

$$\begin{aligned} var\left(\hat{t}_{Y_{AC}^{dip}}\right) &= var\left(\hat{t}_{Y_{GREG}}\right) + \hat{\boldsymbol{\beta}}_M{}^t \; var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \; \hat{\boldsymbol{\beta}}_M \\ &\quad -2 \; \boldsymbol{\beta}^t \; cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M + 2 \; cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M \end{aligned} \tag{2.13}$$

Ricordando la (2.6), possiamo anche scrivere

$$\begin{aligned} var\left(\hat{t}_{Y_{AC}^{dip}}\right) &= var\left(\hat{t}_{Y_{AC}^{dip}}\right) \\ &\quad -2 \; \boldsymbol{\beta}^t \; cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M + 2 \; cov\left(\hat{t}_{Y_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \boldsymbol{\beta}_M \end{aligned}$$

Ceccarelli *e altri* (2013).

Rispetto all'espressione dello stimatore nel caso di indagini indipendenti, nella (2.13) entrano in gioco due ulteriori quantità di cui sarà valutato l'impatto nel paragrafo 2.2.1 e in particolar modo nell'applicazione nel capitolo (4).

Una scrittura equivalente per la (2.13) è data dalla

$$\begin{aligned} var\left(\hat{t}_{Y_{AC}^{dip}}\right) &= \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \left(\gamma_k \; \hat{e}_k\right)\left(\gamma_l \; \hat{e}_l\right) \\ &\quad + \sum_{m=1}^{M} \hat{\beta}_m{}^2 \; var\left(\widetilde{t}_{Z_{1m}}\right) + \sum_{m'\neq m} \hat{\beta}_m \; \hat{\beta}_{m'} \; cov\left(\widetilde{t}_{Z_{1m}}, \widetilde{t}_{Z_{1m'}}\right) \\ &\quad -2 \sum_{q=1}^{M+P}\sum_{m=1}^{M} \hat{\beta}_q \; \hat{\beta}_m \; cov\left(\hat{t}_{U_{HT}}, \widetilde{t}_{Z_{1m}}\right) + 2 \sum_{m=1}^{M} \hat{\beta}_m \; cov\left(\hat{t}_{Y_{HT}}, \hat{t}_{Z_{1m}}\right). \end{aligned}$$

Ceccarelli *e altri* (2013).

Ipotesi di omoschedasticità degli errori

Una valutazione dell'impatto delle informazioni ausiliarie longitudinali sulla varianza della stima di $\hat{t}_{Y_{AC}}$ può essere fornita assumendo l'omoschedasticità degli errori campionari. Inizialmente consideriamo, per semplicità, il caso in cui è stata rilevata sul campione s_2 una sola variabile ausiliaria longitudinale Z, con il totale di controllo $\widetilde{t}_{Z_1}$, quindi con riferimento alla (2.5)

$$Var(A_1) = Var\left(\hat{t}_{Y_{GREG}}\right)$$

$$Var(A_2) = \boldsymbol{\beta}_Z{}^2 \; Var\left(\widetilde{t}_{Z_1}\right)$$

In genere il vincolo longitudinale Z nelle indagini con campioni ruotati è la variabile Y rilevata sul campione s_1 il cui totale, che però continueremo ad indicare con $\widetilde{t}_{Z_1}$, è $\hat{t}_{Y_{t-1}}$. Per questo motivo una semplificazione, adottata generalmente per agevolare la stima della varianza delle differenze nel caso di campioni ruotati e, che riportiamo ora nel nostro caso, è

$$Var\left(\hat{t}_{Y_{GREG}}\right) = Var\left(\widetilde{t}_{Z_1}\right) \tag{2.14}$$

che si basa sull'assunto che il processo di stima (disegno campionario totali di controllo) sia costante tra le due indagini.

In questo modo, sfruttando anche i risultati in Appendice A4, siamo in grado di scrivere:

$$\begin{aligned} Cov\left(\hat{t}_{Y_{GREG}}, \left(t_U - \check{t}_U\right)\beta\right) &= Cov\left(\hat{t}_{Y_{GREG}}, \widetilde{t}_{Z_1}\right)\beta_Z \\ &= r_{YZ}\ \beta_Z\ Var\left(\hat{t}_{Y_{GREG}}\right) \\ &= {r_{YZ}}^2\ Var\left(\hat{t}_{Y_{GREG}}\right), \end{aligned}$$

dove ${r_{XY}}^2$ è il quadrato del coefficiente di regressione tra le stime. Una sua stima può essere data da ${\hat{r}_{YZ}}^2$ calcolato sulla base delle precedenti occasioni di indagine.

Quindi la (2.5) può essere stimata come

$$\begin{aligned} var\left(\hat{t}_{Y_{AC}}\right) &= var\left(\hat{t}_{Y_{GREG}}\right) + {\beta_Z}^2 var\left(\widetilde{t}_{Z_1}\right) - 2\ q(s_c)\ {\hat{r}_{YZ}}^2\ var\left(\hat{\mathbf{t}}_{Y_{GREG}}\right) \\ &= var\left(\hat{t}_{Y_{GREG}}\right)\left(1 - 2\ q(s_c)\ {\hat{r}_{YZ}}^2\right) + {\beta_Z}^2 var\left(\widetilde{t}_{Z_1}\right). \end{aligned} \tag{2.15}$$

Dalla (2.15) abbiamo la dimostrazione che i vincoli longitudinali contribuiscono a ridurre l'errore della stima. Infatti, $2\ q(s_c)\ {\hat{r}_{YZ}}^2$ è una quantità sempre positiva, in quanto $0 \leq {\hat{r}_{YZ}}^2 \leq 1$ e $0 \leq q(s_c) \leq 1$.

Il guadagno in efficienza dipende in maniera diretta dalla correlazione tra la variabili Y e la Z e in maniera inversa dalla proporzione di individui in comune tra s_1 e s_2 in quanto la correlazione tra $\hat{t}_{Y_{GREG}}$ e $\widetilde{t}_{Z_1}$ è nulla sulla parte del campione che non si sovrappone.

Questo risultato può essere facilmente adattato al caso in cui si hanno Z_m variabili ausiliarie longitudinali, con $m = 1, \ldots, M$. È necessario, tuttavia, considerare che l'espressione (2.15) rappresenta un'approssimazione che può portare a valori diversi da quelli che si otterrebbero con la (2.6), in quanto l'ipotesi di omoschedasticità degli errori può discostarsi dalla realtà. Infatti, nelle indagini reali diversi fattori, come ad esempio il meccanismo di mancata risposta e l'*attrition*, possono portare a condizioni diverse da quelle ipotizzate nella (2.14).

Capitolo 3

Ottimalità

Introduzione

Nel capitolo precedente le informazioni ausiliarie campionarie sono state inserite nel processo di stima con la prerogativa di soddisfare la condizione di coerenza esterna. In questo modo sono state determinate le espressioni degli stimatori della varianza ed è stato dimostrato come l'impiego di totali affetti da errore campionario ha un impatto non trascurabile sulle stime. Si è dimostrato, inoltre, come questo impatto viene in parte mitigato quando le informazioni ausiliarie sono delle variabili longitudinali i cui totali di controllo sono stimati su un campione che ha in comune una quota di unità con il campione su cui si basa l'indagine per cui si stanno producendo le stime.

In questo capitolo, invece, l'approccio seguito per inserire le variabili ausiliarie è differente. L'obiettivo, infatti, è di impiegare informazioni ausiliarie con totali di controllo affetti da errore in modo ottimale, condizionatamente alla loro natura campionaria. Seguendo un approccio *design-based* si propone uno stimatore in grado di considerare informazioni ausiliarie con totali di controllo campionari prestando attenzione alla riduzione della varianza.

Questo cambio di prospettiva porta a dei risultati simili ma differenti da quelli ottenuti nel capitolo precedente. In particolar modo, si vedrà come la necessità di minimizzare l'errore campionario porterà a discostarci dalla condizione di coerenza a seconda della "affidabilità" delle stime ottenute su s_1, della relazione tra s_1 e s_2 e della relazione tra le variabili in gioco, ovvero la variabile di interesse Y e i due set di variabili ausiliarie X_p, con $p = 1, \ldots, P$, e Z_m, con $m = 1, \ldots, M$.

Nel paragrafo 3.1 sarà effettuata una breve rassegna della letteratura sul tema

e fornita una rapida descrizione dei metodi. Lo stimatore proposto sarà presentato nel paragrafo 3.2. In particolare nel paragrafo 3.2.1 si illustrerà il modo in cui lo stimatore proposto maneggia le informazioni ausiliarie campionarie e soprattutto i totali di controllo campionari. Il paragrafo 3.2.2, invece, affronta il problema nel suo complesso, ovvero quando ci sono contemporaneamente informazioni ausiliarie con totali di controllo non-campionari e campionari.

In entrambi i paragrafi verranno presentate le espressioni per quattro casi

- $\tilde{\mathbf{t}}_{\mathbf{z}_1}$ è una costante,
- s_1 e s_2 sono indipendenti,
- campioni a due fasi (*two-phase sampling*) $s_2 \subseteq s_1 \subseteq U$,
- s_1 e s_2 campioni complementari, con s_2 estratto nel complemento di s_1, $U \setminus s_1$.

Infine, nel paragrafo 3.2.3, verrà dimostrato sotto quali condizioni con lo stimatore che utilizza totali di controllo campionari si ottiene un guadagno di efficienza rispetto ad altri stimatori.

3.1 Quadro di riferimento

La ricerca di uno stimatore che sfrutta in maniera ottimale le informazioni ausiliarie con totali di controllo stimate da un'altra indagine non è un argomento nuovo. In letteratura sono presenti alcuni lavori sul tema.

Il primo in ordine di tempo è di Zieschang (1986, 1990). Zieschang propone un metodo per ottenere delle stime coerenti per l'indagine sui consumi delle famiglie (Consumer Expenditure (CE) Survey) degli Stati Uniti. La CE divide il campione in due componenti, una sulla quale si rilevano le informazioni tramite intervista e l'altra, la più grande, tramite diario auto-compilato. Le variabili demografiche e diverse informazioni sulle spese vengono rilevate con entrambe le tecniche di rilevazione. Il suo obiettivo è, dunque, quello di migliorare le stime integrando queste informazioni raccolte dalle due componenti. Il metodo che propone, infatti, prevede la determinazione di un doppio sistema di pesi, uno per ciascuna componente, ottenuto aggiustando i pesi da disegno attraverso lo stimatore *GREG*. I pesi aggiustati portano a delle stime, ottenute sulle due componenti prese singolarmente, comparabili tra loro.

Merkouris (2004) propone un'estensione del metodo di Zieschang. In particolare propone l'inserimento nello stimatore di Zieschang di un fattore in grado di tener

conto della differenza nella dimensione effettiva tra due o più indagini. Questa soluzione gli consente di ottenere delle stime migliori e, soprattutto, di ottenere notevoli vantaggi pratici che rendono il metodo facilmente applicabile. Sviluppa, inoltre, l'espressione per stime a livello di dominio (Merkouris, 2010).

Il lavoro centrale in questo contesto, però, è quello di Renssen e Nieuwenbroek (1997) che propongono di utilizzare lo stimatore AR per allineare le stime di due, o più indagini, che hanno delle variabili in comune.

Il metodo da noi proposto ha diversi punti di contatto con il metodo di Renssen e Nieuwembroek. Tuttavia, rispetto a questo e agli altri metodi qui citati, presenta delle differenze sostanziali in quanto considera il processo di stima dell'indagine svolta su s_1 come chiuso, a differenza degli altri che sono applicabili nel caso in cui i processi di stima tra le due indagini si svolgono simultaneamente. Un'altra differenza, che rappresenta un salto metodologico, è dovuta al fatto di sviluppare l'espressione dello stimatore anche nel caso in cui i campioni s_1 e s_2 sono dipendenti e non solo quando sono indipendenti, assunzione che è alla base degli altri metodi.

3.2 Lo stimatore ottimo

L'obiettivo primario in questo contesto è la ricerca di uno stimatore che sia in grado di sfruttare in maniera ottimale le informazioni ausiliarie con totali di controllo stimati da un'altra indagine rinunciando alla condizione di coerenza esterna.

Il contesto di riferimento è sempre quello illustrato nel paragrafo 1.6. Tuttavia, è necessario definire altre quantità come

$$\pi_{1k} = \Pr(k \in s_1) \qquad \pi_{1kl} = \Pr(k, l \in s_1) \quad \Delta_{1kl} = \frac{\pi_{1kl} - \pi_{1k}\ \pi_{1l}}{\pi_{1k}\ \pi_{1l}}$$

$$\pi_{2k} = \Pr(k \in s_2) \qquad \pi_{2kl} = \Pr(k, l \in s_2) \quad \Delta_{2kl} = \frac{\pi_{2kl} - \pi_{2k}\ \pi_{2l}}{\pi_{2k}\ \pi_{2l}}$$

$$\pi_{12kl} = \Pr(k \in s_1, l \in s_2) \qquad \Delta_{12kl} = \frac{\pi_{12kl} - \pi_{1k}\ \pi_{2l}}{\pi_{1k}\ \pi_{2l}}\ .$$

L'elemento di novità rispetto alle quantità definite nel paragrafo 1.6 è π_{12kl}, ovvero la probabilità di inclusione del secondo ordine che ci permette di definire la matrice $\boldsymbol{\Delta}_{12} = (\Delta_{12kl})_{k \in s_1, l \in s_2}$. La matrice $\boldsymbol{\Delta}_{12}$ ci consente di tener conto della relazione tra i campioni s_1 e s_2 e, al contrario delle matrici del disegno $\boldsymbol{\Delta}_1$ e $\boldsymbol{\Delta}_2$, è una matrice non simmetrica $N \times N$. Le relative stime delle matrici da disegno sono

$$\underline{\boldsymbol{\Delta}}_1 = \left(\frac{\Delta_{1kl}}{\pi_{1kl}}\right)_{k,l\in s_1}$$

$$\underline{\boldsymbol{\Delta}}_2 = \left(\frac{\Delta_{2kl}}{\pi_{2kl}}\right)_{k,l\in s_2}$$

$$\underline{\boldsymbol{\Delta}}_2 = \left(\frac{\Delta_{12kl}}{\pi_{12kl}}\right)_{k\in s_1,l\in s_2}$$

Mantenendo sempre la notazione definita nel paragrafo 1.1, definiamo inoltre

$$\begin{aligned}
Var(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}) &= \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X},\\
Var(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}) &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z},\\
Var(\hat{t}_{Y_{HT}}) &= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y},\\
Cov(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}) &= \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z},\\
Cov(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{t}_{Y_{HT}}) &= \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y},\\
Cov(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \hat{t}_{Y_{HT}}) &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}.
\end{aligned}$$

e

$$\begin{aligned}
Var(\tilde{\mathbf{t}}_{\mathbf{Z}_1}) &= \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\\
Cov(\tilde{\mathbf{t}}_{\mathbf{Z}_1}, \hat{\mathbf{t}}_{\mathbf{X}_{HT}}) &= \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{X},\\
Cov(\tilde{\mathbf{t}}_{\mathbf{Z}_1}, \hat{t}_{Y_{HT}}) &= \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}.
\end{aligned}$$

Queste ultime relazioni sono esattamente verificate nel caso in cui le stime $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ siano ottenute con lo stimatore HT, altrimenti rappresentano un'approssimazione la cui differenza con il valore vero diventa trascurabile al crescere della numerosità campionaria di s_1.

Richiamiamo anche lo stimatore AR, già presentato nella (2.4), che considera entrambi i set di variabili ausiliarie che, però, in questo caso scriviamo

$$\hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}^* + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}. \tag{3.1}$$

In maniera equivalente da quanto fatto da (Montanari, 1987) per derivare lo stimatore $GREG_o$, abbiamo ipotizzato che $\boldsymbol{\beta}^*$ e $\boldsymbol{\gamma}$ fossero delle costanti. Questa assunzione, già usata anche nel paragrafo 2.2.2, si basa sulla dimostrazione che la differenza tra $\hat{\boldsymbol{\beta}}^*$ e $\boldsymbol{\beta}^*$ è una quantità di ordine $n^{-1/2}$ che diventa trascurabile al crescere della dimensione del campione (cfr., e.g., Conti e Marella, 2011, p. 88).

Abbiamo quindi derivato l'espressione dello stimatore AR ottimo, AR_o, che è data da

$$\hat{t}_{Y_{AR_o}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}_o^* + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}_o \tag{3.2}$$

in cui $\hat{\boldsymbol{\beta}}_o^*$ e $\hat{\boldsymbol{\gamma}}_o$ sono le stime delle quantità che minimizzano la varianza da disegno dello stimatore AR.

Per spiegare in maniera più chiara il modo in cui i regressori sono determinati e come lo stimatore proposto maneggia i totali di controllo campionari, mostriamo prima il caso in cui abbiamo solo il set di variabili Z e, successivamente, il caso in cui abbiamo entrambi i set di variabili ausiliarie, X e Z.

3.2.1 Totali di controllo non-campionari

Per semplicità partiamo dal caso in cui si considerano solamente variabili ausiliarie con totali noti stimati su s_1 mentre, per il momento, non si considerano variabili ausiliarie con totali di controllo da fonte amministrativa o censuaria. In questo caso lo stimatore espresso nella (3.1) si semplifica in

$$\hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \boldsymbol{\gamma}. \tag{3.3}$$

Seguendo un approccio *design-based*, il valore di $\boldsymbol{\gamma}$ che sostituito nell'espressione precedente consente di minimizzare la varianza da disegno, è dato da[1]

$$\boldsymbol{\gamma}_o = \left[var\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)\right]^{-1} cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}} - \widetilde{\mathbf{t}}_{\mathbf{Z}_1}, \hat{t}_{Y_{HT}}\right),$$

che può anche essere scritto come

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}_2 - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right), \tag{3.4}$$

assumendo che l'inversa esista.

Inserendo la (3.4) nella (3.3) si ottiene lo stimatore AR_o nel caso in cui si usano solo variabili ausiliarie con totali di controllo campionari

$$\hat{t}_{Y_{AR_{oz}}} = \hat{t}_{Y_{HT}} + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \boldsymbol{\gamma}_o. \tag{3.5}$$

Lo stimatore nella (3.5) non è esattamente calibrato su $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$. Infatti, stimando il totale del set di variabili ausiliarie Z si ha

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{Z}_{AR}} &= \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} + \boldsymbol{\Gamma}\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\ &= \boldsymbol{\Gamma}\ \widetilde{\mathbf{t}}_{\mathbf{Z}_1} + (\mathbf{I} - \boldsymbol{\Gamma})\ \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} \end{aligned}$$

dove $\boldsymbol{\Gamma}$ è una matrice $M \times M$ data da

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}_2 - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right), \tag{3.6}$$

[1] Si veda la dimostrazione in Appendice, A5.

e $\mathbf{I}$ è la matrice identità di ordine M.

Dunque, se la fonte da cui sono stimati i totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ è affidabile, cioè se la quantità $\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}$ è piccola in senso assoluto o relativo rispetto a $\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}$, $\boldsymbol{\Gamma}$ si riduce ad una matrice identità. Quando, invece, avviene il contrario $\boldsymbol{\Gamma}$ diventa una matrice di zero. In generale, quindi, lo stimatore nella (3.5) calibra le proprie stime su una quantità tra $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ e $\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}$, ovvero tra la stima ottenuta dall'indagine esterna e la stima derivata dal campione s_2.

Le quantità nella (3.4) e nella (3.6) devono essere stimate e delle loro stime sono ottenute sostituendo le quantità che compongono le espressioni con le relative quantità calcolate sul campione, cioè

$$\hat{\boldsymbol{\gamma}}_o = \left(\mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_1\ \mathbf{Z}_1 + \mathbf{Z}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2 - 2\ \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_{12}\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{y}_2 - \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_{12}\ \mathbf{y}_2\right), \tag{3.7}$$

e

$$\hat{\boldsymbol{\Gamma}} = \left(\mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_1\ \mathbf{Z}_1 + \mathbf{Z}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2 - 2\ \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_{12}\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2 - \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_{12}\ \mathbf{Z}_2\right), \tag{3.8}$$

Nella (3.7) e nella (3.8) compaiono le matrici del disegno $\underline{\boldsymbol{\Delta}}_1$ di dimensione $n_1 \times n_1$, $\underline{\boldsymbol{\Delta}}_2$ di dimensione $n_2 \times n_2$ e $\underline{\boldsymbol{\Delta}}_{12}$ di dimensione $n_1 \times n_2$ stimate sui campioni s_1 e s_2 e le matrici delle variabili ausiliarie rilevate sugli stessi campioni.

Sostituendo $\hat{\boldsymbol{\gamma}}_o$ lo stimatore definito nella (3.5) può anche essere scritto come uno stimatore lineare omogeneo

$$\hat{t}_{AR_{oz}} = \sum_{k \in s_2} w_{ks_2}\ y_k = \mathbf{w}_{AR_{oz}}\ \mathbf{y}_2$$

in cui il vettore dei pesi $\mathbf{w}_{AR_{oz}}$ è uguale a[2]

$$\mathbf{w}_{AR_{oz}} = \mathbf{d}_2 + \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2 \left(\mathbf{Z}_2^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1} \hat{\boldsymbol{\Gamma}} \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right). \tag{3.9}$$

Da questa espressione appare ancora più evidente come la matrice $\hat{\boldsymbol{\Gamma}}$ moduli la condizione di coerenza rispetto ai totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$.

Nel caso in cui i campioni s_1 e s_2 sono estratti mediante un campionamento casuale semplice senza ripetizione è possibile giungere a delle approssimazioni delle espressioni sin qui presentate.

Sia n_C la numerosità dell'intersezione tra i due campioni (cfr. Tam, 1984; Goga, 2008; Qualité e Tillé, 2008, per una descrizione del contesto di riferimento) e

$$\mathbf{C}_\mathbf{z} = \frac{1}{N-1} \left(\mathbf{Z} - \frac{\mathbf{1}\mathbf{1}'\mathbf{Z}}{N}\right)^t \left(\mathbf{Z} - \frac{\mathbf{1}\mathbf{1}'\mathbf{Z}}{N}\right)$$

[2]Si veda dimostrazione in Appendice, A6.

la matrice di varianza e covarianza tra le variabili Z_m, con $m = 1, \ldots, M$, in cui $\mathbf{1}$ è un vettore di uno di dimensione N. È possibile scrivere

$$\begin{aligned}
\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} &= N^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) \mathbf{C}_\mathbf{z}, \\
\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} &= N^2 \left(\frac{1}{n_2} - \frac{1}{N}\right) \mathbf{C}_\mathbf{z}, \\
\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} &= N^2 \left(\frac{\mathbb{E}[n_C]}{n_1\ n_2} - \frac{1}{N}\right) \mathbf{C}_\mathbf{z}.
\end{aligned}$$

Sostituendo queste quantità nella (3.6), si ha che il valore di $\boldsymbol{\Gamma}$ nel caso di campionamento casuale semplice è

$$\begin{aligned}
\widetilde{\boldsymbol{\Gamma}} &= \left\{\left(\frac{1}{n_1} - \frac{1}{N}\right) + \left(\frac{1}{n_2} - \frac{1}{N}\right) - 2\left(\frac{\mathbb{E}[n_C]}{n_1 n_2} - \frac{1}{N}\right)\right\}^{-1} \\
&\quad \times \left\{\left(\frac{1}{n_2} - \frac{1}{N}\right) - \left(\frac{\mathbb{E}[n_C]}{n_1 n_2} - \frac{1}{N}\right)\right\} \mathbf{I} \\
&= \frac{n_1 - \mathbb{E}[n_C]}{n_1 + n_2 - 2\ \mathbb{E}(n_C)} \mathbf{I}. \qquad (3.10)
\end{aligned}$$

Inoltre, se definiamo

$$\mathbf{c_{zy}} = \frac{1}{N-1} \left(\mathbf{Z} - \frac{\mathbf{11'Z}}{N}\right)^\top \left(\mathbf{y} - \frac{\mathbf{11'y}}{N}\right),$$

la (3.4) nel caso di campionamento casuale semplice è

$$\widetilde{\boldsymbol{\gamma}}_o = \frac{n_1 - \mathbb{E}[n_C]}{n_1 + n_2 - 2\ \mathbb{E}[n_C]}\ \mathbf{C}_\mathbf{z}^{-1} \mathbf{c_{zy}}.$$

In questo modo il coefficiente di regressione $\mathbf{C}_\mathbf{z}^{-1}\mathbf{c_{zy}}$ è riscritto in funzione di un coefficiente che tiene conto della sovrapposizione n_C tra i due campioni.

Nei prossimi paragrafi saranno presentati quattro scenari particolari che identificano lo stimatore nella (3.5) come un caso generale di stimatori già noti in letteratura. Gli scenari presentati si riferiscono a situazioni reali che si incontrano di frequente nelle indagini.

Caso 1

Partiamo con il considerare il caso in cui $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ è noto con certezza ed è quindi uguale a $\mathbf{t_Z}$.

Questo comporta che

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} = \underline{\mathbf{0}}$

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \underline{\mathbf{0}}$

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \underline{\mathbf{0}}$

Quindi, si ha

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\right)^{-1} \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

che è equivalente allo stimatore di regressione dello stimatore $GREG_o$ nella (1.18) che utilizza il set di variabili Z. Inoltre $\boldsymbol{\Gamma}$ si riduce a

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^\top \boldsymbol{\Delta}_2 \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{I}$$

e, ovviamente, lo stimatore proposto garantisce il soddisfacimento della condizione di coerenza a dei totali che, in questo caso, sono completamente affidabili. Sostituendo nella (3.3) la stima di $\boldsymbol{\gamma}_o$,

$$\hat{\boldsymbol{\gamma}}_o = \left(\mathbf{Z}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1} \mathbf{Z}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{y}_2.$$

si ottiene esattamente lo stimatore di regressione ottimo di Montanari (1987, pp. 196-197) mostrato nella (1.19).

Questo risultato, pur rappresentando più che altro una verifica del comportamento dello stimatore proposto, è particolarmente importante in quanto rappresenta un ottimo *benchmark* per il nostro stimatore e fa si che questo possa essere visto come un caso generale dello stimatore $GREG_o$.

Caso 2

Nel caso in cui s_1 e s_2 sono estratti in maniera indipendente dalla popolazione U, si ha che

$$\pi_{12kl} = \pi_{1k}\ \pi_{2l}$$

e, di conseguenza, $\boldsymbol{\Delta}_{12} = \mathbf{0}$.
Quindi

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \underline{\mathbf{0}}$

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y} = \underline{\mathbf{0}}$.

Il risultato che si ottiene in questo caso può essere visto come uno stimatore di regressione ridge (cfr. Bardsley e Chambers, 1984; Chambers, 1996; Rao e Singh, 1997) o, più generalmente, come uno stimatore di regressione penalizzata (cfr. Goga e Shezad, 2010; Guggemos e Tillé, 2010). Infatti, $\boldsymbol{\gamma}_o$ si semplifica

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}\right),$$

e può essere visto come il vettore dei coefficienti di una regressione ridge in cui il termine di penalizzazione è uguale a $\boldsymbol{\gamma}_o^t \mathbf{Z}^t \ \boldsymbol{\Delta}_1 \ \mathbf{Z}\boldsymbol{\gamma}_o$, inoltre $\boldsymbol{\Gamma}$

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t \ \boldsymbol{\Delta}_1 \ \mathbf{Z} + \mathbf{Z}^t \ \boldsymbol{\Delta}_2 \ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z}^t \ \boldsymbol{\Delta}_2 \ \mathbf{Z}\right).$$

Nel caso in cui s_1 e s_2 sono due campioni indipendenti estratti con disegno casuale semplice senza ripetizione

$$\mathbb{E}[n_C] = \frac{n_1 \ n_2}{N}$$

e

$$\widetilde{\boldsymbol{\Gamma}} \ = \ \frac{n_1(N-n_2)}{n_1(N-n_2)+n_2(N-n_1)}\mathbf{I}.$$

Quando la dimensione della popolazione N è grande

$$\widetilde{\boldsymbol{\Gamma}} \ = \ \frac{n_1}{n_1+n_2}\mathbf{I}$$

Questo risultato è ottenuto in un contesto analogo anche da Deville (1999, p. 210), Renssen e Nieuwenbroek (1997, p. 371) e da Merkouris (2004, p. 1133).

Caso 3

Assumiamo, invece, che s_2 sia incluso in s_1 $(s_2 \subseteq s_1 \subseteq U)$. Questo caso rappresenta un tipo di disegno campionario chiamato campionamento a due fasi (*two-phase sampling* o anche *double sampling*)[3].

I campioni s_1 e s_2 sono innestati, ovvero da U si estrae il campione s_1 e tra le unità di s_1 si estra il campione s_2 (cfr. Rao, 1973; Cochran, 1977; Hidiroglou e Särndal, 1998). Il campione s_1, dunque, è estratto dalla popolazione U attraverso un disegno campionario basato sulla probabilità di inclusione del primo ordine π_{1k}. Il campione della seconda fase, s_2, è selezionato da s_1 attraverso un disegno campionario con probabilità di inclusione π_{2k}, con $\pi_{2k} = \pi_{k|s_1}$.

Adottando la terminologia di Hidiroglou e Särndal (1998) il nostro caso coincide con quello in cui si considera solo il vettore ridotto delle informazioni ausiliarie, cioè le informazioni ausiliarie sono disponibili a livello del campione s_1, $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$. La stessa situazione è classificata come caso C1 nel lavoro di Estevao e Särndal (2002, p. 236).

Il valore di $\boldsymbol{\gamma}$ che minimizza la varianza del disegno dello stimatore nella (3.3) è

$$\boldsymbol{\gamma}_o \ = \ \left(\mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_1}, \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]\right)^{-1} \left(\mathbb{E}\left[Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \hat{t}_{Y_{HT}}\right)|s_1\right]\right).$$

[3]Questo tipo di campionamento è utilizzato ad esempio nelle indagini Multiscopo Istat per condurre dei focus su particolari argomenti o sottopopolazioni, come avviene ad esempio per le disabilità a partire dal campione dell'indagine sulle condizioni di salute e ricorso ai servizi sanitari, oppure sulle discriminazioni a partire dall'indagine sulla sicurezza dei cittadini.

Poiché nel campionamento a due fasi

- $Cov\left(\tilde{\mathbf{t}}_{\mathbf{Z}_1}, \hat{t}_{Y_{HT}}\right) = \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{y}$
- $Cov\left(\tilde{\mathbf{t}}_{\mathbf{Z}_1}, \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) = \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}$
- $Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \hat{t}_{Y_{HT}}\right) = \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{y} + \mathbb{E}\left[Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \hat{t}_{Y_{HT}}\right)|s_1\right]$
- $Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) = \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]$.

$$\boldsymbol{\gamma}_o = \left(\mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]\right)^{-1}\left(\mathbb{E}\left[Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \hat{t}_{Y_{HT}}\right)|s_1\right]\right)$$

e

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]\right)^{-1}\left(\mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]\right) = \mathbf{I}.$$

Lo stimatore, quindi, nel caso di campioni innestati garantisce la coerenza con $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$.

Caso 4

Consideriamo ora il caso in cui s_2 è selezionato in $U \setminus s_1$, cioè nel complemento di s_1. Per questo motivo i campioni s_1 e s_2 sono detti *complementary samples* o anche *negative samples* perché le covarianze calcolate tra i due campioni sono negative (cfr. Ardilly e Tillé, 2006, p. 35-38). Questo situazione si riferisce al caso in cui si desidera effettuare una rotazione rigida tra i due campioni, in modo che s_1 e s_2 non abbiano unità in comune.

Questa strategia di campionamento viene generalmente utilizzata per minimizzare il carico sui rispondenti (*respondent burden*).

Quando $s_1 \cap s_2 = \emptyset$, se indichiamo con I_{k1} e I_{k2} la variabile indicatrice della presenza dell'unità k in s_1 o s_2

$$\begin{aligned}\Delta_{12kl} &= Cov\left(\frac{I_{1k}}{\pi_{1k}}, \frac{I_{2l}}{\pi_{2l}}\right)\\ &= \mathbb{E}\left[Cov\left(\frac{I_{1k}}{\pi_{1k}}, \frac{I_{2l}}{\pi_{2l}}\middle| s_1\right)\right] + Cov\left(\mathbb{E}\left[\frac{I_{1k}}{\pi_{1k}}|s_1\right], \mathbb{E}\left[\frac{I_{2l}}{\pi_{2l}}|s_1\right]\right).\end{aligned}$$

Il primo termine è nullo e

$$\mathbb{E}\left(\frac{I_{2l}}{\pi_{2l}}|s_1\right) = \frac{1 - I_{1l}}{1 - \pi_{1l}} = \frac{1 - I_{1l}}{\pi_{1l}}\frac{\pi_{1l}}{1 - \pi_{1l}}$$

(Guandalini e Tillé, 2014).

Così, se definiamo $\mathbf{D}$ la matrice diagonale $N \times N$ che contiene l'elemento $\pi_{1l}/(1 - \pi_{1l})$ sulla sua diagonale si ha che

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = -\mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{Z}$

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y} = -\mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{y}$

quindi

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{Z}\right)^{-1}\left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{y}\right).$$

Inoltre,

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{Z}\right)^{-1}\left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\mathbf{D}\ \mathbf{Z}\right).$$

Nel caso di estrazione con campionamento casuale semplice di s_1 e s_2, $n_C = 0$ e

$$\widetilde{\boldsymbol{\Gamma}} = \frac{n_1}{n_1 + n_2}\mathbf{I}.$$

3.2.2 Totali di controllo non-campionari e campionari

Passiamo ora al caso in cui si hanno a disposizioni entrambi i set di variabili ausiliarie X e Z. In questo caso quindi consideriamo lo stimatore AR, già presentato nella (3.1),

$$\hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}^* + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}.$$

Tuttavia, invece che determinare l'espressione della varianza da disegno di questa espressione, consideriamo l'espressione equivalente

$$\hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}} - \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\mathbf{L}}\hat{\boldsymbol{\gamma}} + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}, \qquad (3.11)$$

in cui $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^* - \hat{\mathbf{L}}\hat{\boldsymbol{\gamma}}$ e $\hat{\mathbf{L}} = \left(\mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{X}_2\right)^{-1}\mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2$ è la stima della matrice dei coefficienti di regressione del set di variabili Z sullo spazio generato dalle colonne della matrice $\mathbf{X}$.

Quest'espressione è la scrittura "ortogonalizzata" dell'espressione nella (3.1), in cui i vettori $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ sono ortogonali tra loro (cfr. Seber, 1977). Come vedremo, questa scrittura ci consente di sfruttare i risultati dello stimatore $GREG_o$ di Montanari, illustrato nel paragrafo 1.3. Considerando $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ costanti, infatti, i valori di questi vettori che minimizzano la varianza da disegno dello stimatore AR sono[4]

$$\boldsymbol{\beta}_o = \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

[4]Si veda dimostrazione in Appendice, A7.

che è il coefficiente di regressione che minimizza la varianza da disegno dello stimatore $GREG$, già illustrato nella (1.18) e

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right) \tag{3.12}$$

con

$$\begin{aligned} \mathbf{R} &= \boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_2\ \mathbf{X} \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2 \\ &= \boldsymbol{\Delta}_2\ (\mathbf{I} - \mathbf{P}) \end{aligned}$$

dove $\mathbf{P}$ è la matrice degli operatori di proiezione sullo spazio generato dalle colonne della matrice $\mathbf{X}$.

La struttura di $\boldsymbol{\gamma}_o$ è dovuta in parte alla relazione tra regressori parziali nel caso in cui sono aggiunti ulteriori regressori nel modello (Seber, 1977, p. 54) e in parte all'errore campionario, che può essere interpretato come errore di misura, che li affligge (Fuller, 1987, p. 105).

Sostituendo le rispettive stime di $\boldsymbol{\beta}_o$ e di $\boldsymbol{\gamma}_o$ nella (3.11) si ha

$$\hat{t}_{Y_{AR_o}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}_o - \left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\mathbf{L}}\hat{\boldsymbol{\gamma}}_o + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}_o. \tag{3.13}$$

Lo stimatore AR_o soddisfa la condizione di coerenza esterna con i totali di controllo non-campionari, infatti, quando si stimano i totali delle X_p, con $p = 1, \ldots, P$

$$\mathbf{B} = \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X} = \mathbf{I}$$

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{X} - \mathbf{X}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{X}\right) = \mathbf{0}$$

in quanto $\mathbf{Z}^t\ \mathbf{R}\ \mathbf{X} = \mathbf{0}$, poiché $\mathbf{X}$ e $\mathbf{RZ}$ sono ortogonali ($\mathbf{X} \perp \mathbf{RZ}$) e, $\mathbf{X}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{X} = \mathbf{0}$, dalla prima equazione normale. Quindi

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{X}_{AR_o}} &= \hat{\mathbf{t}}_{\mathbf{X}_{HT}} + \mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}} \\ &= \mathbf{t}_\mathbf{X} \end{aligned}$$

Quando, invece, si stimano i totali delle Z_m, con $m = 1, \ldots, M$ variabili ausiliarie

$$\mathbf{B} = \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{L}$$

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right). \tag{3.14}$$

Quindi

$$\hat{\mathbf{t}}_{\mathbf{Z}_{AR_o}} = \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} + \left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\mathbf{L}} - \left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\mathbf{L}}\hat{\boldsymbol{\Gamma}} + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\Gamma}}.$$

La condizione di coerenza rispetto ai totali campionari può essere vista in maniera più chiara guardando al sistema di pesi dello stimatore AR_o

$$\mathbf{w}_{AR_o} = \mathbf{w}_{CAL} + \hat{\mathbf{R}}\ \mathbf{Z}_2 \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right)^{-1} \hat{\mathbf{\Gamma}} \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{Z}_2{}^t\ \mathbf{w}_{CAL}\right). \tag{3.15}$$

Come avviene per il sistema di pesi dello stimatore AC nella (2.3), il soddisfacimento della condizione di coerenza con i totali di controllo campionari stimati su s_1 viene ottenuto come residuale dopo che è stata soddisfatta la condizione di coerenza rispetto ai totali di controllo non campionari, $\mathbf{t}_{\mathbf{X}}$. La differenza principale tra le due espressioni è dovuta alla presenza $\hat{\mathbf{\Gamma}}$.

Nella (3.15) emerge il ruolo della matrice $\hat{\mathbf{\Gamma}}$ che opera da fattore di *shrinkage* allentando o stringendo la condizione di coerenza sui totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$. Dalla (3.14), si può vedere come $\mathbf{\Gamma}$ e conseguentemente una sua stima, sia funzione della

- varianza dei totali noti campionari, $\mathbf{Z}^t\ \mathbf{\Delta}_1\ \mathbf{Z}$;
- relazione che lega i campioni s_1 e s_2;
- relazione che lega il set di variabili Z con il set di variabili X, $\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}$.

Il suo ruolo nella determinazione dello stimatore AR_o e, soprattutto, nella minimizzazione della varianza è cruciale. Se i totali di controllo campionario hanno una varianza troppo grande, $\mathbf{\Gamma}$ tende a diventare una matrice di 0 e la condizione di coerenza su $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ verrà via via allentata, poiché, anche in relazione ai risultati visti nel capitolo precedente, si introdurrebbe una quantità di errore eccessiva e quindi, in questo caso, l'impiego di informazioni ausiliarie campionarie sarebbe inefficiente. Quando, invece, i totali di controllo campionari provengono da una fonte altamente affidabile, o le stime delle due indagini sono altamente correlate per via del loro disegno campionario, $\mathbf{\Gamma}$ stringe la condizione di coerenza fino, come vedremo in seguito, a diventare esattamente $\mathbf{I}$ nel caso in cui i totali sono noti con certezza (sono totali di controllo non-campionari) o nel caso del campionamento a due fasi.

Quando $\mathbf{\Gamma}$ tende alla matrice identità lo stimatore AR_o tenderà allo stimatore $GREG_o$ che considera entrambi i set di variabli. La condizione di coerenza sarà, dunque, soddisfatta sia per le X che per le Z. Quando, invece, $\mathbf{\Gamma}$ sarà uguale a $\underline{\mathbf{0}}$, il set di variabili ausiliarie Z non sarà incluso nel processo di stima, ma queste saranno considerate come variabili *extra*.

Come fatto nel paragrafo 3.2.1 consideriamo il caso di campioni estratti attraverso un disegno casuale semplice senza ripetizione. Questa semplificazione consentirà, anche in questo contesto, di cogliere ulteriori aspetti dello stimatore proposto. Definiamo, quindi, oltre a $\mathbf{C}_{\mathbf{Z}}$

$$\mathbf{C_x} = \tfrac{1}{N-1}\left(\mathbf{X} - \tfrac{\mathbf{11'X}}{N}\right)^t \left(\mathbf{X} - \tfrac{\mathbf{11'X}}{N}\right)$$

$$\mathbf{C_{xz}} = \tfrac{1}{N-1}\left(\mathbf{X} - \tfrac{\mathbf{11}^t\mathbf{X}}{N}\right)^t \left(\mathbf{Z} - \tfrac{\mathbf{11'Z}}{N}\right) = \mathbf{C}^t_{\mathbf{zx}}$$

la matrice di varianza e covarianza $P \times P$ tra le variabili X e la matrice di covarianze $P \times M$, e rispettivamente $M \times P$ tra le variabili X e le Z. Quindi

$$\mathbf{X}^t\ \mathbf{\Delta}_2\ \mathbf{X} = N^2\left(\tfrac{1}{n_2} - \tfrac{1}{N}\right)\mathbf{C_x}$$

$$\mathbf{X}^t\ \mathbf{\Delta}_{21}\ \mathbf{Z} = N^2\left(\tfrac{\mathbb{E}[n_C]}{n_1\ n_2} - \tfrac{1}{N}\right)\mathbf{C_{xz}}$$

$$\mathbf{Z}^t\ \mathbf{\Delta}_{12}\ \mathbf{X} = N^2\left(\tfrac{\mathbb{E}[n_C]}{n_1\ n_2} - \tfrac{1}{N}\right)\mathbf{C_{zx}}$$

$$\mathbf{X}^t\ \mathbf{\Delta}_2\ \mathbf{Z} = N^2\left(\tfrac{1}{n_2} - \tfrac{1}{N}\right)\mathbf{C_{xz}}$$

$$\mathbf{Z}^t\ \mathbf{\Delta}_2\ \mathbf{X} = N^2\left(\tfrac{1}{n_2} - \tfrac{1}{N}\right)\mathbf{C_{zx}}.$$

Inoltre

$$\mathbf{X}^t\ \mathbf{\Delta}_2\ \mathbf{y} = N^2\left(\tfrac{1}{n_2} - \tfrac{1}{N}\right)\mathbf{c_{xy}},$$

dove

$$\mathbf{c_{xy}} = \frac{1}{N-1}\left(\mathbf{X} - \frac{\mathbf{11'X}}{N}\right)^t \left(\mathbf{y} - \frac{\mathbf{11'y}}{N}\right).$$

Utilizzando queste espressioni nella (3.14), ovvero nel caso in cui si vogliono stimare i totali $\mathbf{t_Z}$, si ha che

$$\begin{aligned}\widetilde{\mathbf{\Gamma}} \;=\; & \left\{[n_1 + n_2 - 2\ \mathbb{E}[n_C]]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C_z^{-1}C_{zx}C_x^{-1}C_{xz}}\right\}^{-1} \\ & \times \left\{[n_1 - \mathbb{E}[n_C]]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C_z^{-1}C_{zx}C_x^{-1}C_{xz}}\right\} \qquad (3.16)\end{aligned}$$

Questa espressione può essere vista come una generalizzazione della espressione (3.10). Infatti, nella (3.16) oltre alla dimensione dell'errore delle stime compare un fattore di attenuazione sia nel "numeratore" che nel "denominatore" di $\widetilde{\mathbf{\Gamma}}$.

Il fattore di attenuazione è proporzionale alla matrice $\mathbf{C_z^{-1}C_{zx}C_x^{-1}C_{xz}}$. Nell'analisi delle correlazioni canoniche (cfr., e.g., Mardia *e altri*, 1979; Vitali, 1993; Härdle e Simar, 2011) la matrice $\mathbf{C_z^{-1}C_{zx}C_x^{-1}C_{xz}}$ è la matrice i cui autovalori, i coefficienti di correlazione canonica, misurano la correlazione tra due gruppi di variabili. Possiamo, dunque, affermare che, in termini geometrici, dipende dalla sovrapposizione

degli spazi generati rispettivamente dalle colonne della matrice $\mathbf{X}$ e della matrice $\mathbf{Z}$. Tradotto in termini pratici, dipende dalla capacità del set di variabili Z di fornirci informazioni *nuove* per la variabile di interesse non fornite, in parte o del tutto, dal set di variabili X.

Questo risultato, fornisce un'importate indicazione pratica nella scelta delle informazioni ausiliarie da inserire nel processo di stima. Queste, infatti, oltre a essere delle stime affidabili, devono apportare un contributo ulteriore rispetto alle variabili già considerate.

Ripercorriamo, così come fatto per lo stimatore AR_o con i soli totali di controllo campionari, i quattro casi particolari quando si hanno a disposizione entrambi i set di variabili ausiliarie.

Caso 1

Nel caso 1 consideriamo $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ noti con certezza e quindi uguali a $\mathbf{t}_{\mathbf{Z}}$. Questo comporta che

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} = \underline{\mathbf{0}}$
- $\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \underline{\mathbf{0}}$

Quindi, si ha

$$\boldsymbol{\beta}_o = \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$
$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\right)^{-1} \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

ovvero il coefficiente di regressione delle X che minimizza la varianza da disegno dello stimatore $GREG$, già illustrato nella (1.18), è uguale al coefficiente di regressione delle Z. Sostituendo le rispettive stime

$$\hat{\boldsymbol{\beta}}_o = \left({\mathbf{X}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{X}_2\right)^{-1} {\mathbf{X}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{y}_2$$
$$\hat{\boldsymbol{\gamma}}_o = \left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1} {\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{y}_2,$$

nella (3.1) si ottiene ancora un'espressione equivalente a quella dello stimatore $GREG_o$ nel caso, però, in cui vengono utilizzati i due set di variabili. Le X e le Z hanno lo stesso ruolo nella stima e la condizione di coerenza con i totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ è soddisfatta. Infatti, quando si stima $\mathbf{t}_Z$, $\mathbf{B} = \mathbf{L}$ e $\boldsymbol{\Gamma} = \mathbf{I}$ e sostituendo le relative stime nella (3.1)

$$\begin{aligned}
\hat{\mathbf{t}}_{\mathbf{Z}_{AR}} &= \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} + \hat{\mathbf{L}}^t \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) - \hat{\mathbf{L}}^t \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) + \mathbf{I}\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\
&= \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} + \widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}} \\
&= \widetilde{\mathbf{t}}_{\mathbf{Z}_1}.
\end{aligned}$$

Caso 2

Quando s_1 e s_2 sono assunti essere indipendenti, $\mathbf{\Delta}_{12} = \underline{\mathbf{0}}$. Mentre l'espressione dello stimatore $\boldsymbol{\beta}_o$, e della relativa stima, rimangono le stesse già presentate,

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \mathbf{\Delta}_1\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y}\right)$$

e la sua stima

$$\hat{\boldsymbol{\gamma}}_o = \left(\mathbf{Z_2}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2 + \mathbf{Z}_1^t\ \mathbf{\Delta}_1\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z_2}^t\ \hat{\mathbf{R}}\ \mathbf{y}_2\right).$$

Anche quando vengono considerati contemporaneamente i due set di variabili ausiliarie $\boldsymbol{\gamma}_o$ può essere visto come vettore di coefficienti di una regressione ridge in cui il termine di penalizzazione è uguale a $\boldsymbol{\gamma}_o^t \mathbf{Z}^t\ \mathbf{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma}_o$.

Tuttavia, a differenza del caso equivalente nel paragrafo 3.2.1, la regressione non è tra la variabile di interesse e le variabili Z_m, con $m = 1, \ldots, M$, ma tra la variabile di interesse e i residui della regressione delle Z_m rispetto alle variabili X_p, con $p = 1, \ldots, P$.

In questo caso la condizione di coerenza con i totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ non è soddisfatta in quanto

$$\mathbf{\Gamma} = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \mathbf{\Delta}_1\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}\right)$$

e la relativa stima

$$\hat{\mathbf{\Gamma}} = \left(\mathbf{Z_2}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2 + \mathbf{Z}_1^t\ \mathbf{\Delta}_1\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z_2}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right).$$

sono diverse dalla matrice identità e dipendono dalla varianza dei residui della regressione del set di variabili Z rispetto al set di variabili X e dalla varianza delle stime $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$.

Nel caso di campionamento casuale semplice

$$\begin{aligned}
\widetilde{\mathbf{\Gamma}} &= \left\{\left[n_1 + n_2 - 2\ \frac{n_1\ n_2}{N}\right]\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right\}^{-1} \\
&\quad \times \left\{\left[n_1 - \frac{n_1\ n_2}{N}\right]\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right\} \\
&= \left\{[n_1(N-n_2) + n_2(N-n_1)]\,\mathbf{I} - [n_1(N-n_2)]\,\mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right\}^{-1} \\
&\quad \times \left\{[n_1(N-n_2)]\left[\mathbf{I} - \mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right]\right\}.
\end{aligned}$$

Quando la dimensione della popolazione è grande

$$\begin{aligned}
\widetilde{\mathbf{\Gamma}} &= \left\{[n_1 + n_2]\,\mathbf{I} - n_1\mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right\}^{-1} \\
&\quad \times \left\{n_1\left[\mathbf{I} - \mathbf{C}_\mathbf{z}^{-1}\mathbf{C}_\mathbf{zx}\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{xz}\right]\right\}.
\end{aligned}$$

.

Caso 3

Quando s_2 è incluso in s_1 ($s_2 \subset s_1 \subset U$), oltre alle relazioni illustrate nel paragrafo 3.2.1, abbiamo

$$Var\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) = \mathbf{X}^t\ \boldsymbol{\Delta}_1\ \mathbf{X} + \mathbb{E}\left[Var\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)|s_1\right]$$

$$Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \widetilde{\mathbf{t}}_{\mathbf{Z}_1}\right) = \mathbf{X}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}$$

$$Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) = \mathbf{X}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbb{E}\left[Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)|s_1\right]$$

$$Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{t}_{Y_{HT}}\right) = \mathbf{X}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + \mathbb{E}\left[Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \hat{\mathbf{t}}_{\mathbf{y}_{HT}}\right)|s_1\right].$$

Ricordiamo, inoltre, che

$$\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} = \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}$$

$$\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y} = \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

Anche in questo caso, con i campioni innestati, lo stimatore ottimo soddisfa la condizione di coerenza rispetto ai totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$, infatti sostituendo queste espressioni nella (3.14) si ottiene la matrice identità.

Nel caso di un campionamento casuale semplice, $\mathbb{E}[n_C] = n_2$, quindi la (3.16) si riduce a

$$\begin{aligned}
\widetilde{\boldsymbol{\Gamma}} &= \left\{[n_1 + n_2 - 2n_2]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\}^{-1} \\
&\quad \times \left\{[n_1 - n_2]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\} \\
&= \left\{[n_1 - n_2]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\}^{-1} \\
&\quad \times \left\{[n_1 - n_2]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right]\mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\} = \mathbf{I}.
\end{aligned}$$

Caso 4

Nell'ultimo caso, consideriamo il campione s_2 estratto nel complemento di s_1, $U \setminus s_1$. Come dimostrato nel paragrafo 3.2.1 nel caso di campioni complementari, possiamo scrivere

$$\boldsymbol{\Delta}_{12} = -\boldsymbol{\Delta}_1\mathbf{D},$$

dove la matrice $\mathbf{D}$ è una matrice diagonale $N \times N$ che contiene l'elemento $\pi_{1l}/(1-\pi_{1l})$ sulla sua diagonale, quindi

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} + 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_1 \mathbf{D}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \hat{\mathbf{R}}\ \mathbf{y} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1 \mathbf{D}\ \mathbf{y}\right)$$

Indicando con $\hat{\mathbf{D}}$ la matrice $\mathbf{D}$ stimata sul campione s_1 di dimensione $n_1 \times n_2$ e sostituendola nelle espressioni otteniamo una stima di $\boldsymbol{\gamma}_o$ data da

$$\hat{\boldsymbol{\gamma}}_o = \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2 + \mathbf{Z}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z} + 2\ \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_1 \hat{\mathbf{D}}\ \mathbf{Z}_2\right)^{-1} \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{y}_2 + \mathbf{Z}_1{}^t\ \underline{\boldsymbol{\Delta}}_1 \hat{\mathbf{D}}\ \mathbf{y}_2\right).$$

Le stime prodotte con lo stimatore AR_o in questo caso non sono coerenti con i totali noti campionari in quanto

$$\boldsymbol{\Gamma} = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_1 \mathbf{D}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_1 \mathbf{D}\ \mathbf{Z}\right).$$

Se, invece, si considera il caso di campionamento casuale semplice

$$\begin{aligned} \widetilde{\boldsymbol{\Gamma}} &= \left\{[n_1 + n_2]\,\mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right] \mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\}^{-1} \\ &\quad \times \left\{n_1\ \mathbf{I} - \left[\frac{n_1(N-n_2)}{N}\right] \mathbf{C}_{\mathbf{z}}^{-1}\mathbf{C}_{\mathbf{zx}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{xz}}\right\}. \end{aligned}$$

poichè $\mathbb{E}\left[n_C\right] = 0$.

3.2.3 Effetto stimatore

Abbiamo sottolineato più volte nel corso del capitolo come lo stimatore proposto, AR_o, regolando attraverso $\boldsymbol{\Gamma}$ la condizione di coerenza con i totali $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$, sia in grado di minimizzare la varianza da disegno.

In questo paragrafo, vogliamo confrontare la varianza da disegno dello stimatore AR_o con quella di altri stimatori per vedere in quali casi l'impiego delle informazioni ausiliarie campionarie è inefficiente.

La varianza da disegno dello stimatore AR_o è[5]

$$Var\left(\hat{t}_{Y_{AR_o}}\right) = \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o \quad (3.17)$$

Dalla (3.17) si ha che si ottiene sicuramente un guadagno di efficienza rispetto allo stimatore HT. Infatti, dalla varianza da disegno dello stimatore HT (riportata nella (1.8)), $\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$, vengono sottratte due quantità positive in quanto

$$\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o = \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

[5]Si veda la dimostrazione in Appendice, A8.

$$\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\boldsymbol{\gamma}_o = \left(\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)\boldsymbol{\gamma}_o$$
$$= \left(\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)\left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1}\left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right)$$

sono due forme quadratiche semi-definite positive. Il risultato di ottenere un guadagno anche se il set di variabili Z ha totali di controllo campionari è dovuto, come detto più volte, alla struttura del vettore $\boldsymbol{\gamma}$ che tiene conto di questo aspetto.

Il guadagno di efficienza che si ottiene con lo stimatore AR_o, però è stemperato dalla quantità $\mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o$ che tiene conto, come avevamo già visto e sottolineato nel paragrafo 3.2.2, del grado di sovrapposizione tra le informazioni ausiliarie Z_m e X_p.

L'effetto stimatore (*deft*) di AR_o, calcolato come il rapporto tra la varianza da disegno dello stimatore AR_o e lo stimatore HT, ci conferma questo risultato, infatti

$$deft\left(\hat{t}_{Y_{AR_o}}\right) = \frac{Var\left(\hat{t}_{Y_{AR_o}}\right)}{Var\left(\hat{t}_{Y_{HT}}\right)} = \frac{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o}{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}}$$

Questa quantità è sicuramente inferiore ad 1 in quanto, come già detto prima $\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o$ è una forma quadratica semi-definita positiva. Lo stimatore AR_o, quindi, è sempre più conveniente dello stimatore HT ed il suo guadagno è pari a

$$\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o.$$

Confrontiamo ora, invece, la varianza da disegno dello stimatore AR_o con lo stimatore $GREG_o$ illustrato nel paragrafo 1.4.1 che considera solamente il set di variabili ausiliarie X, quindi con i totali di controllo non-campionari. Considerando la varianza da disegno dello stimatore $GREG_o$ nella (1.17),

$$\frac{Var\left(\hat{t}_{Y_{AR_o}}\right)}{Var\left(\hat{t}_{Y_{GREG_o}}\right)} = \frac{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o}{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o}$$

Anche in questo caso questa quantità è sempre inferiore ad uno ma, come era lecito attendersi, il guadagno rispetto allo stimatore $GREG_o$ è minore rispetto a quello che si ottiene rispetto allo stimatore HT, infatti si riduce a

$$\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}.$$

in quanto è relativo al solo guadagno ottenuto con il set di variabili Z.

Concludiamo il confronto con gli altri stimatori, con il caso in cui lo stimatore $GREG_o$ considera entrambi i set di variabili ausiliarie. La varianza da disegno dello stimatore $GREG_o$ in questo caso sarà

$$\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o$$

dove si fa notare, il valore di $\boldsymbol{\gamma}$ che minimizza la varianza da disegno, indicato con $\bar{\boldsymbol{\gamma}}_o$ per evitare confusione, è

$$\bar{\boldsymbol{\gamma}}_o = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}\right)^{-1} \mathbf{Z}^t\ \mathbf{R}\ \mathbf{y}.$$

mentre per lo stimatore AR_o è

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right).$$

Il rapporto tra i due stimatori, in questo caso è

$$\frac{Var\left(\hat{t}_{Y_{AR_o}}\right)}{Var\left(\hat{t}_{Y_{GREG_o}}\right)} = \frac{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o}{\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta}_o - \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\bar{\boldsymbol{\gamma}}_o}$$

Si ottiene che la varianza da disegno dello stimatore AR_o è al massimo uguale alla varianza dello stimatore $GREG_o$, cioè si ha che

$$\frac{Var\left(\hat{t}_{Y_{AR_o}}\right)}{Var\left(\hat{t}_{Y_{GREG_o}}\right)} \geq 1,$$

in quanto

$$\frac{\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma}_o - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}_o}{\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\bar{\boldsymbol{\gamma}}_o} \geq 1$$

Questo risultato può essere spiegato in due modi. Il primo si ricollega a i risultati ottenuti nel capitolo precedente e da Berger *e altri* (2009) e Dever e Valliant (2010), ovvero che nel caso in cui si utilizzano informazioni ausiliarie campionarie e si calcola la varianza dello stimatore come se i totali non fossero affetti da errore campionario si sottostima il reale errore delle stime. L'errore di sottostima è pari a

$$\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\left(\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}}_o\right) - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}$$

Il secondo motivo, invece, ci fornisce un'indicazione pratica, anche un po' ovvia. Ovvero se si hanno a disposizione totali di controllo non affetti da errore è sempre più conveniente usare quelli.

Uno stimatore della varianza da disegno di AR_o è dato da

$$var\left(\hat{t}_{Y_{AR}}\right) = \mathbf{y_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{y}_2 - \mathbf{y_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{X}_2\hat{\boldsymbol{\beta}}_o - \mathbf{y_2}^t\ \hat{\mathbf{R}}\ \mathbf{Z}\hat{\boldsymbol{\gamma}}_o + \mathbf{y_2}^t\ \underline{\boldsymbol{\Delta}}_{21}\ \mathbf{Z}_2\hat{\boldsymbol{\gamma}}_o \tag{3.18}$$

ovvero sostituendo nella (3.17) le opportune stime campionarie.

Un inconveniente degli stimatori costruiti come abbiamo fatto per lo stimatore AR_o proposto è che in alcuni casi possono portare a stime instabili (cfr. e.g. Montanari, 1988; Rao, 1994; Merkouris, 2004). Questo accade perlopiù al crescere del numero delle informazioni ausiliarie, ovvero quando il numero di gradi di libertà dell'errore, $n - (P + M) - 1$ diventa piccolo.

In generale, quindi è preferibile non abusare con il numero di informazioni ausiliarie da inserire nel processo di stima o in alternativa ricorrere a delle approssimazioni, come ad esempio utilizzando le espressioni relative al caso di campionamento casuale semplice senza ripetizione che sono state presentate.

Capitolo 4

Studio di simulazione ed applicazioni

Introduzione

Nei due capitoli precedenti abbiamo presentato due stimatori che sono in grado di inserire nel processo di stima informazioni ausiliarie con totali di controllo affetti da errore campionario. I due stimatori seguono due differenti obiettivi, lo stimatore AC ricerca la coerenza anche a discapito dell'efficienza delle stime, mentre lo stimatore AR_o, allentando in maniera opportuna questo vincolo, porta a delle stime con un errore minore rispetto a quello degli stimatori che utilizzano lo stesso set di variabili ausiliarie.

Nel primo paragrafo di questo capitolo saranno riportati i risultati relativi ad uno studio di simulazione svolto per confrontare le distribuzioni campionarie di questi due stimatori con altri stimatori illustrati nei precedenti capitoli. In particolare saranno confrontati con lo stimatore HT, con lo stimatore $GREG_o$, sia basato solo sul set di variabili X che su entrambi i set di variabili X e Z, con lo stimatore CAL, ovviamente basato sul solo set di variabili X, e con lo stimatore RW.

La simulazione è svolta in modo da mettere in luce come la correlazione tra le variabili in gioco influenza le perfomance degli stimatori ed evidenziare i contesti di applicazione più favorevoli per ciascuno di questi.

Per comprendere più a fondo lo stimatore AC, inoltre, nel secondo paragrafo di questo capitolo saranno riportate le stime e i relative errori campionari ricavati applicando lo stimatore ai dati dell'indagine sui redditi e le condizioni di vita (It-Silc, *Italian Survey on Income and Living Conditions*) e sui dati mensili della rilevazione

italiana sulle Forze di Lavoro (RCFL, Rilevazione Continua delle Forze di Lavoro).

Queste due indagini rappresentano due casi reali in cui si fa ricorso in maniera sistematica ad informazioni ausiliarie di tipo campionario. L'indagine It-Silc, infatti, utilizza le stime relative alla condizione lavorativa e al livello di istruzione calcolate dalla RCFL per migliorare le stime sui redditi. La RCFL, invece, ci consente di applicare lo stimatore AC nel caso sviluppato nel paragrafo 2.2.2, in quanto nel processo di stima vengono utilizzate informazioni ausiliarie longitudinali.

I risultati dello studio di simulazione e le applicazioni ci consentiranno di apprezzare in maniera più chiara le proprietà degli stimatori proposti e, soprattutto, ci forniranno delle indicazioni per un loro uso più consapevole.

4.1 Studio di simulazione

L'obiettivo dello studio di simulazione è stato quello di confrontare le distribuzioni campionarie degli stimatori proposti con altri stimatori noti. In particolar modo sono stati confrontati

- stimatore di Horvitz-Thompson, HT (cfr. (1.4) e (1.6))
- stimatore di regressione generalizzata ottimo, $GREG_o$ (cfr. (1.19) e (1.17))
  - che considera il solo set di variabili X, $GREG_{o_x}$
  - che considera entrambi i set di variabili X e Z, $GREG_{o_{xz}}$
- stimatore calibrato, CAL (cfr. (1.20) e (1.24))
- stimatore Repeated-Weighting, RW (cfr. paragrafo 2.1)
- stimatore AC (cfr. (2.2) e (2.6) e (2.13))
- stimatore AR (cfr. (3.2) e (3.17))

La popolazione di interesse è stata definita generando $N = 10.000$ vettori $(y, x, z)^t$ da una distribuzione Normale multivariata

$$(y, x, z)^t \sim MN\left(\boldsymbol{\mu}, \sigma^2 \mathsf{P}\right)$$

con $\boldsymbol{\mu} = (20, 20, 20)^t$, $\sigma^2 = 10$ e matrice di correlazione

$$\mathsf{P} = \begin{pmatrix} 1 & r_{yx} & r_{yz} \\ r_{xy} & 1 & r_{xz} \\ r_{zy} & r_{zx} & 1 \end{pmatrix}.$$

La matrice di correlazione ci consente di prendere in considerazione diverse possibili combinazioni delle correlazioni tra le variabili per comprendere meglio e approfondire il comportamento degli stimatori proposti.

Definendo in maniera diversa la matrice P abbiamo riprodotto un quadro esaustivo delle diverse combinazioni delle correlazioni parziali tra le variabili. Classificando il quadrato del coefficiente delle correlazioni parziali in tre livelli

- Alto (H), $0,75 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 1,00$
- Medio (M), $0,40 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 0,60$
- Basso (L), $0,00 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 0,25$

abbiamo, infatti 27 possibili incroci tra le tre correlazioni parziali che possiamo calcolare sulle variabili Y, X e Z. A partire dai coefficienti di regressione semplici abbiamo determinato la matrice P in modo da riprodurre una popolazione una popolazione in cui le variabili Y, X e Z presentassero le correlazioni parziali richieste (cfr. Tabella 4.1 per una rappresentazione schematica, invece, Appendice. A9 per i risultati di tutti gli scenari implementati).

Tabella 4.1: Scenari dello studio di simulazione.

Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$	Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$	Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$
1	H	H	H	10	M	H	H	19	L	H	H
2	H	H	M	11	M	H	M	20	L	H	M
3	H	H	L	12	M	H	L	21	L	H	L
4	H	M	H	13	M	M	H	22	L	M	H
5	H	M	M	14	M	M	M	23	L	M	M
6	H	M	L	15	M	M	L	24	L	M	L
7	H	L	H	16	M	L	H	25	L	L	H
8	H	L	M	17	M	L	M	26	L	L	M
9	H	L	L	18	M	L	L	27	L	L	L

Per ciascuno scenario è stata, dunque, generata una popolazione definendo in maniera opportuna la matrice di correlazione P. Da questa popolazione sono stati selezionati con campionamento casuale semplice senza ripetizione i campioni s_1 e s_2 di numerosità rispettivamente $n_1 = 400$ e $n_2 = 200$. I campioni, inoltre sono stati estratti in modo tale da riprodurre i quattro casi già visti nel paragrafo 3.2, ovvero

- $\widetilde{\mathbf{t}}_{\mathbf{Z}_1}$ costante;
- s_1 e s_2 campioni indipendenti;
- campioni a due fasi (*two-phase sampling*), $s_2 \subseteq s_1 \subseteq U$;
- s_1 e s_2 campioni complementari, con s_2 estratto nel complemento di s_1, $U \setminus s_1$.

Sul campione s_2, quindi, sono state calcolate le stime del totale della Y con tutti gli stimatori elencati sopra. Il totale di controllo della variabile X è stato ottenuto sommando i valori della variabile nella popolazione, mentre quello della variabile Z, $\widetilde{t}_{Z_1}$, è stato stimato di volta in volta sul campione s_1 con lo stimatore HT. Questa procedure è stata replicata 1.000 volte per comparare la distribuzione campionaria degli stimatori.

Come misure di sintesi sono stati calcolati l'*empirical Relative Bias*

$$RB\left(\hat{t}_{Y\cdot}\right) = \frac{\sum_{i=1}^{1.000}\left(\hat{t}_{Y\cdot i} - t_Y\right)}{t_Y}$$

ed il *relative Root Mean Squared Error*

$$rRMSE(\hat{t}_Y) = \frac{\sqrt{\sum_{i=1}^{1.000}\left(\hat{t}_{Y\cdot i} - t_Y\right)^2}}{t_Y}.$$

con $i = 1, \ldots, 1.000$ numero delle replicazioni.

Nella Tabella 4.2 è presentato il caso in cui nella popolazione la correlazione tra le variabili è rappresentata dalla matrice

$$\mathsf{P} = \begin{pmatrix} 1 & 0,878 & -0,213 \\ 0,878 & 1 & 0,139 \\ -0,213 & 0,139 & 1 \end{pmatrix}.$$

che riproduce lo scenario 5 (H M M), ovvero le correlazioni parziali nella popolazione sono

- ${r_{(yx|z)}}^2 = 0,879$;
- ${r_{(yz|x)}}^2 = 0,497$;
- ${r_{(xz|y)}}^2 = 0.484$.

Dalla Tabella 4.2 e 4.3 ([4] e [5]) possiamo vedere come tutti gli stimatori considerati, tranne lo stimatore HT che non ha nessun controllo sulle variabili ausiliarie, sono calibrati su t_X. Infatti, forniscono sempre il valore di t_X.

Il caso 1 rappresenta un'eccezione per la condizione di coerenza con $\tilde{t}_{Z_1}$ che in quel caso è uguale a t_Z (colonne [6] e [7]). Lo stimatore $GREG_{o_{xz}}$ e lo stimatore AC solo gli unici che soddisfano sempre la condizione di coerenza con $\tilde{t}_Z$ mentre lo stimatore AR_o, oltre al caso 1, la soddisfa solamente nel caso 3.

Tabella 4.2: Stime di t_Y, t_X and t_Z, RB%, rRMSE% ottenute con gli stimatori HT, $GREG_{o_x}$, $GREG_{o_{xz}}$, RW, AC e AR, per diverse relazioni tra i campioni s_1, con $n1 = 400$, e $s2$, con $n2 = 200$, estratti con campionamento casuale semplice senza ripetizione da una popolazione di dimensione N=10.000. Il livello della correlazione tra le variabili è modulato dalla matrice P. Caso 1 e Caso 2.

Simatore	$\mu(\hat{t}_y)$ [1]	RBias% [2]	RrMSE% [3]	$\mu(\hat{t}_x)$ [4]	$sd(\hat{t}_x)$ [5]	$\mu(\hat{t}_z)$ [6]	$sd(\hat{t}_z)$ [7]
			Caso 1				
$\hat{t}_X$)=	200.091,963			$\widetilde{t}_{Z_1} = t_Z$=200.468,236			
HT	200.089,764	- 1,099	35,593	200.148,961	2.249,687	200.427,158	2.309,112
$GREG_{o_x}$	200.126,164	17,093	17,357	200.189,813	0,000	200.486,421	2.888,467
$GREG_{o_{xz}}$	200.118,965	13,495	11,878	200.189,813	0,000	200.468,236	0,000
CAL	200.133,412	20,715	17,824	200.189,813	0,000	200.486,421	2.888,467
RW	200.136,200	22,108	16,385	200.189,813	0,000	200.468,236	0,000
AC	200.136,200	22,108	16,385	200.189,813	0,000	200.468,236	0,000
AR_o	200.118,965	13,495	11,878	200.189,813	0,000	200.468,236	0,000
			Caso 2				
$\mu(\widetilde{t}_{z_1})$=	200.422,538			$sd(\widetilde{t}_{z_1})$=1.538,333			
HT	200.066,046	-12,953	35,802	200.122,520	2.257,655	200.410,145	2.207,523
$GREG_{o_x}$	200.125,394	16,708	16,554	200.189,813	0,000	200.496,755	2.927,279
$GREG_{o_{xz}}$	200.124,273	16,148	14,354	200.189,813	0,000	200.422,538	1.538,333
CAL	200.135,689	21,853	16,987	200.189,813	0,000	200.496,755	2.927,279
RW	220.142,821	25,418	16,469	200.189,813	0,000	200.431,360	1.678,521
AC	200.142,184	25,099	16,419	200.189,813	0,000	200.422,538	1.538,333
AR_o	200.124,817	16,419	13,555	200.189,813	0,000	200.419,538	1.274,746

Tabella 4.3: Stime di t_Y, t_X and t_Z, RB%, rRMSE% ottenute con gli stimatori HT, $GREG_{o_x}$, $GREG_{o_{xz}}$, RW, AC e AR, per diverse relazioni tra i campioni s_1, con $n1 = 400$, e $s2$, con $n2 = 200$, estratti con campionamento casuale semplice senza ripetizione da una popolazione di dimensione N=10.000. Il livello della correlazione tra le variabili è modulato dalla matrice P. Caso 3 e Caso 4.

Simatore	$\mu(\hat{t}_y)$ [1]	RBias% [2]	RrMSE% [3]	$\mu(\hat{t}_x)$ [4]	$sd(\hat{t}_x)$ [5]	$\mu(\hat{t}_z)$ [6]	$sd(\hat{t}_z)$ [7]
				Caso 3			
	$\mu(\widetilde{t}_{z_1})$= 200.368.260				$sd(\widetilde{t}_{z_1})$=2.233,485		
HT	200.075,148	- 8,404	35,485	200.142,920	2.268,548	200.368,260	2.233,485
$GREG_{o_x}$	200.112,887	10,457	16,591	200.189,813	0,000	200.434,017	2.888,069
$GREG_{o_{xz}}$	200.115,531	11,779	16,694	200.189,813	0,000	200.368,260	2.233,485
CAL	200.124,720	16,371	17,167	200.189,813	0,000	200.434,017	2.888,069
RW	200.124,720	16.371	17,167	200.189,813	0,000	200.434,017	2.888,069
AC	200.123,065	15,544	18,941	200.189,813	0,000	200.368,260	2.233,485
AR_o	200.115,531	11,779	16,694	200.189,813	0,000	200.368,260	2.233,485
				Caso 4			
	$\mu(\widetilde{t}_{z_1})$= 200.520,096				$sd(\widetilde{t}_{z_1})$=2,248.208		
HT	200.059,570	-16,189	34,749	200.182,989	2.231,619	200.539,410	2.189,489
$GREG_{o_x}$	200.064,825	-13,536	16,723	200.189,813	0,000	200.565,652	2.847,492
$GREG_{o_{xz}}$	200.072,528	- 9,713	17,393	200.189,813	0,000	200.520,096	2.339,075
CAL	200.069,455	-11,248	17,322	200.189,813	0,000	200.565,652	2.847,492
RW	200.068,309	-11,821	16,694	200.189,813	0,000	200.543,694	2.058,885
AC	200.068,873	-11,540	16,942	200.189,813	0,000	200.520,096	2.339,075
AR_o	200.068,223	-11,864	14,687	200.189,813	0,000	200.531,800	1.579,034

Proprio in corrispondenza del caso 3, lo stimatore AR coincide con lo stimatore $GREG_{o_{xz}}$. Questo caso rappresenta una particolarità in quanto è l'unico caso in cui la condizione di coerenza esterna sui totali noti campionari e la condizione di ottimalità sono soddifatte entrambe. Si fa notare che pur riportando agli stessi totali di Z, lo stimatore AC e lo stimatore AR non coincidono in quanto si basano su due sistemi di pesi differenti (cfr. (2.3) e (3.15)). In generale lo stimatore AR_o è in grado implicitamente di minimizzare la variabilità delle stime t_Z [7] e questo si traduce in una minor variabilità di $\hat{t}_Y$.

Le stime prodotte con AR_o presentano una distorsione minore rispetto a quelle ottenute con gli altri stimatori [2]. Lo stimatore RW in alcuni casi può portare ad una distorsione che differisce molto rispetto a quella che si ottiene con lo stimatore AC e con lo stimatore AR_o.

Anche in termini di efficienza lo stimatore AR svolge un ruolo centrale rispetto agli altri indipendentemente dallo scenario considerato.

Lo seguono a ruota lo stimatore AC e lo stimatore RW che hanno delle prestazioni molto simili. I due stimatori consentono di ottenere un maggiore guadagno di efficienza nel caso in cui la correlazione parziale tra la X e la Z è bassa. In caso contrario portano a delle stime inefficienti, in quanto l'errore che viene importato nella stima di t_Y non è compensato da una migliore definizione del modello di regressione. Lo stesso avviene nel caso in cui le Z sono poco correlate con la Y, l'efficienza degli stimatori che le considerano - $GREG_{o_{xz}}$, RW e AC - peggiora notevolmente fino, in alcuni casi, a renderli meno efficienti dello stimatore $GREG_{o_x}$ che non le considera o, in casi estremi, anche dello stimatore HT.

Nel caso in cui la correlazione che lega le variabili X e Z con la Y ha segno diverso, o nel caso in cui la variabile X ha una correlazione bassa con la Y, lo stimatore RW risulta meno efficiente dello stimatore AR.

Fa eccezione lo stimatore AR che implicitamente rilascia il vincolo di coerenza, γ_o sarà prossimo allo 0, e, proprio in questi casi, si dimostra particolarmente favorevole rispetto agli altri stimatori.

Lo stimatore AR_o sembra, quindi, inequivocabilmente, preferibile rispetto agli altri e soprattutto rispetto ai suoi diretti competitor, RW e AC. Tuttavia, guardare a questi stimatori come in contrapposizione tra loro è sbagliato. Rappresentano, infatti, metodi per inserire nel processo di stima informazioni ausiliarie campionarie che tengono conto di necessità diverse, ovvero la condizione di coerenza esterna con i totali noti campionari. Qualora questa fosse richiesta, lo stimatore AR_o non potrebbe essere impiegato in quanto non gode di questa proprietà, almeno che non

ci si trovi nel caso di campioni innestati, mentre può essere usato lo stimatore AC, o anche lo stimatore RW, che, come abbiamo dimostrato, forniscono comunque delle stime efficienti a patto che il loro impiego sia consapevole dell'impatto che l'errore dei totali di controllo campionari ha sulle stime.

A tal proposito nel prossimo paragrafo verranno illustrate due applicazioni dello stimatore AC che metteranno in evidenza proprio questo aspetto.

4.2 Applicazioni a dati reali

Nel paragrafo precedente abbiamo presentato lo studio di simulazione effettuato per confrontare lo stimatore AC e lo stimatore AR_o con altri stimatori. Lo stimatore AR_o è risultato particolarmente affidabile in quanto consente di ottenere stime efficienti indipendentemente dal livello di correlazione tra le variabili ausiliarie che si vogliono considerare e la variabile di interesse. Più attenzione, invece, necessità l'impiego dello stimatore AC, che però, rispetto ad AR_o, soddisfa la condizione di coerenza.

Di seguito presenteremo i risultati ottenuti applicando lo stimatore AC nel processo di stima di due indagini che utilizzano totali di controllo campionari per migliorare le stime e per limitare il fenomeno di mancata risposta e sotto-copertura.

Le due indagini rappresentano esattamente i due casi tipo per l'applicazione delle espressioni della varianza dello stimatore AC illustrate nel paragrafo 2.2. Nel paragrafo 4.2.1 verrà applicata l'espressione nella (2.6), ovvero relativa allo stimatore della varianza di AC nel caso di totali di controllo campionari da un'indagine indipendente. Nell'indagine It-Silc, infatti, per migliorare le stime di categorie di reddito particolarmente ostili alla rilevazioni vengono utilizzati totali noti campionari sulla condizione lavorativa e sul livello di istruzione stimati dalla RCFL.

Nel paragrafo 4.2.2, invece, verrà applicata l'espressione dello stimatore della varianza di AC nel caso di totali di controllo campionari da un'indagine dipendente (2.13). Nella RCFL, infatti, viene adottato un disegno campionario con uno schema di rotazione del campione. Per migliorare le stime vengono utilizzate variabili ausiliarie longitudinali con totali campionari ottenuti considerando una quota consistente di unità in comune con quelle della rilevazione in corso.

Nelle applicazioni verrà messa in risalto la sottostima dell'errore quando si assume che l'errore associato ai totali di controllo campionari ha un impatto trascurabile sull'errore delle stime prodotte e può essere ignorato e, soprattutto, come nel caso dello stimatore AC la valutazione dell'introduzione di informazioni ausiliarie deve

essere svolta in maniera approfondita ed attenta.

4.2.1 La rilevazione sui redditi e sulle condizioni di vita

L'indagine sui redditi e sulle condizioni di vita (It-Silc) è una rilevazione annuale e fornisce stime su reddito, povertà ed esclusione sociale[1]. È armonizzata a livello europeo e deve rispettare determinati standard qualitativi nel processo di produzione dei dati.

I campioni annuali, composti da circa 32.000 famiglie e 60.000 individui, sono estratti mediante un disegno campionario a due stadi in cui le unità di primo stadio, i comuni, sono stratificati per regione e dimensione demografica mentre le unità di secondo stadio, le famiglie, sono selezionate dai registri di popolazione dei comuni estratti.

Eurostat suggerisce di utilizzare uno stimatore di tipo CAL e di calibrare le stime a totali di controllo noti da fonte amministrativa, ovvero popolazione residente per regione, sesso ed età e numero di famiglie residenti. L'Istat ha seguito ed ampliato questa direttiva. Attualmente, infatti, vengono considerati nel processo di stima 163 vincoli di cui 141 noti da fonte amministrativa (variabili demografiche desunte dal Censimento, Bilancio demografico, Bilancio demografico degli stranieri) e 22 provenienti dalla RCFL del quarto trimestre dell'anno di riferimento del reddito.

I vincoli campionari utilizzati sono relativi alla condizione lavorativa e al livello di istruzione. Sono stati inseriti per cercare di migliorare le stime di categorie particolarmente ostili alla rilevazione sui redditi, come quella dei lavoratori autonomi.

Con riferimento alla Tabella 4.4, sono riportate le stime dei totali dei redditi per diverse categorie e ripartizioni territoriali (NUTS I) su un campione del 2008 relativo a 20.928 famiglie e 52.433 individui. I totali di controllo campionari delle variabili sulla condizione lavorativa e il livello di istruzioni sono stati stimati sul campione del quarto trimestre del 2007 della RCFL.

In Tabella 4.4 sono riportate due stime dei totali dei redditi, una [1] ottenuta considerando i 141 totali di controllo noti da fonti amministrative, ovvero come se avessimo usato lo stimatore CAL (1.20) e l'altra [3] considerando, invece, tutti i 163 totali di controllo, quindi, anche i 22 totali di controllo campionari dalla RCFL ottenuta con lo stimatore AC (2.2).

[1] Per approfondire l'argomento si rimanda a Istat (2008)

Tabella 4.4: Stime di varie tipologie di reddito per ripartizione territoriale e relativi errori campionari ottenuti con lo stimatore CAL e con lo stimatore AC, dati It-Silc 2008 e RCFL quarto trimestre 2007.

	no totali RCFL		**totali RCFL**		
	Stima		Stima	non random	random
Livello NUTS I	del totale (in milioni di €) [1]	CV% [2]	del totale (in milioni €) [3]	CV% [4]	CV% [5]
			Reddito familiare		
Nord-Est	222.792	1,229	222.573	1,010	1,106
Nord-ovest	156.728	1,044	156.737	0,965	1,087
Centro	154.183	1,213	153.750	1,026	1,143
Sud	131.922	1,240	129.954	0,960	1,143
Isole	63.042	1,748	62.484	1,267	1,565
ITALIA	728.667	0,575	725.497	0,478	0,538
			Reddito individuale		
Nord-Est	221.674	1,127	221.455	0,910	1,027
Nord-Ovest	155.520	1,042	155.538	0,898	1,038
Centro	152.886	1,227	152.452	1,009	1,156
Sud	130.682	1,227	128.714	0,875	1,084
Isole	62.334	1,811	61.757	1,241	1,544
ITALIA	723.096	0,557	719.916	0,444	0,515
			Reddito da lavoro		
Nord-Est	147.170	1,546	146.627	1,246	1,481
Nord-Ovest	104.429	1,448	104.151	1,218	1,355
Centro	101.509	1,714	101.047	1,360	1,416
Sud	84.233	1,762	81.825	1,279	1,431
Isole	38.708	3,061	37.853	1,843	2,067
ITALIA	476.049	0,789	471.503	0,615	0,695
			Reddito da lavoro autonomo		
Nord-Est	43.440	5,024	43.405	3,724	4,252
Nord-Ovset	30.554	4,858	30.612	3,640	4,308
Centro	30.193	4,725	30.312	3,777	4,464
Sud	21.684	4,876	21.077	4,063	4,790
Isole	9.246	6,780	9.066	6,013	7,219
ITALIA	135.119	2,399	134.472	1,851	2,158
			Reddito da lavoro dipendente		
Nord-Est	103.730	1,345	103.221	0,922	1,334
Nord-Ovest	73.874	1,413	73.539	0,878	1,411
Centro	71.316	1,756	70.735	1,065	1,635
Sud	62.549	1,904	60.748	1,052	1,689
Isole	29.460	3,483	28.788	1,939	2,620
ITALIA	340.930	0,780	337.031	0,479	0,723
			Reddito da pensione		
Nord-Est	59.985	1,344	60.274	1,259	1,381
Nord-Ovest	39.473	1,340	39.749	1,280	1,490
Centro	42.347	1,549	42.391	1,442	1,621
Sud	37.426	1,507	37.835	1,337	1,531
Isole	19.079	2,124	19.305	1,875	2,131
ITALIA	198.310	0,684	199.554	0,633	0,712

Le stime ottenute con i due diversi set di informazioni ausiliarie differiscono leggermente tra loro, tuttavia le conclusioni più interessanti sono quelle che si ricavano analizzando gli errori delle stime calcolati con il coefficiente di variazione percentuale

$$CV\% = \frac{\sqrt{var\left(\hat{t}_{Y.}\right)}}{\hat{t}_{Y.}}\ 100$$

che ci fornisce una misura relativa dell'errore campionario e ci consente di fare dei confronti.

In particolar modo in Tabella 4.4 sono riportati il

- $CV\%$ dello stimatore CAL che considera solo i 141 totali di controllo non-campionari calcolato con lo stimatore della varianza di CAL nella (1.24) [2],
- $CV\%$ dello stimatore AC che considera tutti i 163 totali di controllo non-campionari e campionari calcolato però con lo stimatore della varianza di CAL nella (1.24), quindi ignorando l'errore importato dai totali noti campionari [4],
- $CV\%$ dello stimatore AC che considera tutti i 163 totali di controllo non-campionari e campionari, calcolato con lo stimatore della varianza di AC nel caso di indagini indipendenti nella (2.6) che tiene conto dell'errore dovuto all'impiego di totali noti campionari [5].

Confrontando la [4] e la [5] si vede immediatamente come ipotizzare che i totali di controllo campionari non abbiano impatto sull'errore delle stime porta una sottostima del reale errore. La differenza tra queste due quantità è dovuta al secondo elemento nella (2.6) che, rispetto alla (1.24), tiene conto dell'aumento di errore. Tuttavia, bisogna sottolineare come questo aumento sia consistente solo in alcuni casi, quelli dovuti ad aggregati più piccoli come le isole, poiché si sono considerati dei totali di controllo che pur essendo campionari risultano essere molto affidabili (hanno un errore piccolo).

Se, invece, confrontiamo la [5] con la [2] abbiamo dei risultati contraddittori. Si ha un grandissimo guadagno di efficienza per le stime del reddito da lavoro autonomo che rappresenta sempre una categoria particolarmente difficile nelle indagini sui redditi, mentre si peggiorano leggermente le stime del reddito da pensione. Leggeri miglioramenti si verificano, invece, per tutte le altre categorie.

In alcuni domini il ricorso ai totali di controllo campionari porta a stime più efficienti in altri, invece, ne peggiora l'efficienza. Questo consente di mettere in evidenza un aspetto importante che può essere dovuto principalmente alla correlazione tra le variabili ausiliarie campionarie e la variabile di interesse. Ad esempio la

condizione occupazione e titolo di studio non sono particolarmente correlate con i redditi da pensione e quindi la riduzione del primo membro nella (2.6) non è tale da compensare l'aumento dovuto all'uso di totali di controllo campionari stimati dalla RCFL.

In generale, inoltre, quando si aggiungono informazioni ausiliarie campionarie, il guadagno di efficienza è maggiore nel caso in cui l'errore di partenza era maggiore e minore quando era minore (Merkouris, 2010; Traat e Särndal, 2011), come ad esempio avviene rispettivamente per il reddito da lavoro autonomo ed il reddito familiare.

Per l'indagine It-Silc, nonostante l'aumento dell'errore nei domini delle stime del reddito da pensione, l'introduzione di totali campionari porta notevoli miglioramenti nell'efficienza delle stime e quindi è auspicabile.

4.2.2 La rilevazione continua sulle Forze di Lavoro

La rilevazione continua sulle forze di lavoro (RCFL) è una rilevazione continua effettuata tutte le settimane dell'anno. Rappresenta la fonte ufficiale del mercato del lavoro in Italia e fornisce principalmente le stime di occupati, disoccupati e non forze lavoro[2].

Il disegno campionario è a base trimestrale. Ciascun campione trimestrale è estratto attraverso un disegno campionario a due stadi (comuni-famiglie) con stratificazione delle unità di primo stadio per provincia e dimensione demografica mentre le unità di secondo stadio, le famiglie, sono selezionate dai registri di popolazione dei comuni estratti.

Il disegno campionario prevede uno schema di rotazione del campione di tipo $(2_T, 2_T, 2_T)$. Le famiglie estratte rimangono nel campione per due trimestri consecutivi, dopo una pausa di due trimestri rientrano nel campione per altri due trimestri e successivamente escono definitivamente dalla rilevazione.

L'Istat, in accordo con Eurostat, ha previsto anche uno schema di rotazione nel tempo per fornire delle stime mensili e garantire, quindi, un'adeguata rappresentatività dei campioni mensili. Poiché però i campioni mensili hanno un terzo della numerosità dei campioni trimestrali, l'Istat ha deciso di inserire delle informazioni ausiliarie longitudinali rilevate tre mesi prima (un trimestre) e dodici mesi prima (quattro trimestri) per ottenere delle stime delle serie meno volatili attraverso dei campioni più stabili e con una struttura più simile tra loro.

[2]Per approfondire l'argomento si rimanda a Istat (2006)

Figura 4.1: Schema di rotazione della rilevazione continua sulle Forze di Lavoro, RCFL (Istat, 2006, p. 41).

Anno	Gruppi di rotazione										
I trimestre anno a	A4	B3			E2	F1					
II trimestre anno a		B4	C3			F2	G1				
III trimestre anno a			C4	D3			G2	H1			
IV trimestre anno a				D4	E3			H2	I1		
I trimestre anno a+1					E4	F3			I2	J1	
II trimestre anno a+1						F4	G3			J2	K1

Per lo schema di rotazione (riportato in Figura 4.1), la sovrapposizione dei campioni ad un trimestre di distanza e a quattro trimestri di distanza, dovrebbe essere del 50%, ma in realtà a causa della mancata risposta e soprattutto dell'*attrition* è di circa il 30% per entrambi.

Il sistema di informazioni ausiliarie nella RCFL è particolarmente complesso e prevede 302 variabili ausiliarie, di cui 206 con totali noti da fonte amministrativa (variabili demografiche desunte dal Bilancio demografico, Stima Rapida e Bilancio demografico degli stranieri) e 96 totali di controllo campionari relativi alle stime di occupati, disoccupati e non forze lavoro di tre mesi prima e di dodici mesi prima.

Nella Tabella 4.5 abbiamo seguito lo stesso schema adottato nell'applicazione sui dati dell'It-Silc. Dunque, oltre alle stime ottenute con lo stimatore CAL [1] e con lo stimatore AC [3] sono riportati

- $CV\%$ dello stimatore CAL che considera solo i 206 totali di controllo non-campionari calcolato con lo stimatore della varianza di CAL nella (1.24) [2],
- $CV\%$ dello stimatore AC che considera tutti i 302 totali di controllo non-campionari e campionari calcolato però con lo stimatore della varianza di CAL nella (1.24), quindi ignorando l'errore importato dai totali noti campionari [4],
- $CV\%$ dello stimatore AC che considera tutti i 302 totali di controllo non-campionari e campionari, calcolato con lo stimatore della varianza di AC (2.6) quindi come se le indagini fossero indipendenti [5].
- $CV\%$ dello stimatore AC che considera tutti i 306 totali di controllo non-campionari e campionari, calcolato con lo stimatore della varianza di AC nel caso di indagini dipendenti nella (2.13) che tiene conto dell'errore dovuto all'impiego di totali noti campionari e del grado di sovrapposizione tra i campioni [6].

Tabella 4.5: Stime di occupati, disoccupati e non forze lavoro e relativi errori campionari ottenuti con lo stimatore CAL e con lo stimatore AC, dati RCFL Settembre 2009.

	no totali RCFL		**totali RCFL**			
				non random	random	
	Stima del totale [1]	CV% [2]	Stima del totale [3]	CV% [4]	ind. CV% [5]	dip. CV% [6]
Italia						
Occupati	22.786.251	0,440	22.886.373	0,331	0,414	0,403
Disoccupati	2.021.889	2,678	2.031.044	2,260	2,727	2,727
Non Forze Lavoro	34.982.481	0,275	34.873.584	0,216	0,255	0,251
Maschi						
Occupati	13.599.617	0,491	13.647.567	0,385	0,467	0,448
Disoccupati	1.092.265	3,508	1.093.438	3,094	3,157	3,157
Non Forze Lavoro	14.371.974	0,435	14.323.056	0,341	0,400	0,397
Femmine						
Occupati	9.186.634	0,828	9.238.806	0,610	0,754	0,742
Disoccupati	926.624	3,926	937.606	3,259	3,864	3,864
Non Forze Lavoro	20.610.507	0,360	20.550.531	0,275	0,323	0,317

L'inserimento delle informazioni ausiliarie campionarie porta un cambiamento nelle stime ottenute in tutti i domini di stima ([1] vs. [3]). Concentrandoci sui coefficienti di variazione percentuale, rispetto al caso dell'It-Silc, abbiamo delle variazioni importanti.

Si può notare, infatti, che l'impiego di informazioni ausiliarie longitudinali migliora l'efficienza delle stime. Il miglioramento è dovuto all'elevata correlazione tra le informazioni ausiliarie longitudinali e la variabile di interesse grazie alla dipendenza tra i due campioni. L'elevata correlazione tra le informazioni ausiliarie longitudinali e la variabile di interesse contribuisce a diminuire il primo termine della (2.13), come si può notare confrontando [2] e [4]. Anche la dipendenza tra i due campioni, come dimostrato nel paragrafo 2.2.2, contribuisce a diminuire l'errore della stima. Una misura del guadagno che si ha considerando variabili ausiliarie longitudinali è data dal confronto tra [5] e [6].

Confrontando, invece, le differenze in termini relativi della [4] e la [5] nella Tabella 4.5 con le quantità omologhe nella Tabella 4.4 si può notare come la scelta della fonte da cui prendere le stime come totali di controllo ha un impatto non trascurabile sulle stime. In questo caso, ricorrendo a delle stime ottenute su un campione

mensile che in proporzione è un terzo di quello da cui sono prese le stime utilizzati come totali di controllo campionari in It-Silc, fa sì che il secondo elemento della (2.6), o equivalentemente della (2.13), sia maggiore. Le stime ottenute su i campioni mensili presentano inevitabilmente un errore maggiore in quanto hanno una numerosità campionaria di gran lunga inferiore a quelle delle stime trimestrali. Ad ogni modo, anche in questo caso l'uso di informazioni ausiliarie campionarie fatto in maniera mirata porta notevoli vantaggi nella riduzione dell'errore delle stime.

Conclusioni

Il ricorso ad informazioni ausiliarie campionarie è un argomento di particolare importanza ed attualità per quel che riguarda il processo di stima nelle rilevazioni campionarie.

L'introduzione di informazioni ausiliarie attraverso il ricorso a totali noti da fonte amministrativa è una prassi largamente diffusa e consolidata. Tuttavia, è sempre maggiore, e lo sarà ancor di più, l'esigenza di introdurre informazioni ausiliarie i cui totali sono noti da fonte campionaria (quindi affetti da errore).

L'uso di queste informazioni risulta particolarmente utile per diversi motivi, su tutti la possibilità di inserire nel processo di stima informazioni ausiliarie particolarmente importanti per la variabile di interesse che altrimenti non si potrebbero considerare in quanto non sono disponibili totali da fonte amministrativa. I vantaggi che derivano, dunque, dall'uso di informazioni ausiliarie campionarie possono essere collocati su due livelli. In questo contesto, infatti, le proprietà che si possono richiedere allo stimatore che ricorre a informazioni ausiliarie di tipo campionario sono la coerenza esterna e l'ottimalità. La ricerca del soddisfacimento di queste due proprietà, che sono tra loro mutuamente esclusive, ha portato ad alcuni importanti risultati.

Nel caso in cui la prerogativa principale è quella di ottenere delle stime coerenti con stime prodotte da altre indagini, prerogativa che può essere molto importante per gli Istituti nazionali di statistica, è stato proposto lo stimatore AC. Lo stimatore AC può anche essere visto come un'estensione dello stimatore CAL che consente di soddisfare contemporaneamente il vincolo di coerenza con totali di controllo sia non-campionari che campionari.

Quando si ricorre allo stimatore AC, però, assume particolare importanza la valutazione dell'impatto dell'errore importato nelle stime considerando totali di controllo campionari, quindi affetti loro stessi da errore. A tal proposito è stata proposta un'espressione dello stimatore della varianza campionaria di AC. Tale espressione è stata sviluppata tenendo conto della relazione che intercorre tra il campione su

cui si vogliono produrre le stime della variabile di interesse e quello, o quelli, da cui sono desunte le stime utilizzate come totali di controllo (campionari).

Sono state, infatti, presentati due stimatori per la stima della varianza campionaria di AC che rispondono al caso in cui i due campioni sono indipendenti e al caso in cui sono dipendenti, ovvero si ricorre ad informazioni ausiliarie longitudinali.

Le espressioni proposte dimostrano come l'efficienza delle stime prodotte con lo stimatore AC dipenda dalla correlazione tra la variabile di interesse e le variabili ausiliarie, l'errore dei totali di controllo campionari e la dipendenza tra i campioni.

Nel caso in cui, invece, non è necessario soddisfare la condizione di coerenza è stato proposto lo stimatore AR_o. Questo stimatore consente di ottenere delle stime che hanno varianza da disegno minima rispetto agli altri stimatori che utilizzano le stesse informazioni ausiliarie. Infatti, attraverso un fattore di *shrinkage*, è in grado di allentare o restringere la condizione di coerenza con i totali campionari in modo da minimizzare la varianza delle stime prodotte. Determinando attraverso un approccio *design-based* i coefficienti di regressione si è derivato uno stimatore in grado di valutare l'errore della stima che si otterrebbe utilizzando lo stimatore di HT, l'errore dei totali campionari e il grado di dipendenza tra le due indagini in modo da rendere minima la varianza delle stime.

I due stimatori proposti confrontati anche con altri stimatori tramite uno studio di simulazione presentano buone proprietà. In particolar modo lo stimatore AR_o risulta particolarmente vantaggioso, in quanto, è in grado implicitamente di valutare l'affidabilità dei totali di controllo utilizzati e di minimizzare la varianza. Il suo vantaggio rispetto agli altri è maggiore soprattutto quando le informazioni ausiliarie inserite sono poco correlate con la variabile di interesse.

Poichè l'errore delle stime prodotte con lo stimatore AC, invece, dipende principalmente dalla correlazione che lega le variabili in gioco e dall'errore delle stime utilizzate come totali di controllo, il suo uso deve essere valutato in maniera più attenta. Infatti, applicando lo stimatore AC, e quindi informazioni ausiliarie campionarie, su dati reali si possono ottenere delle stime migliori in alcuni domini, ma leggermente peggiori in altri. In questo caso, dunque, il ruolo del ricercatore come conoscitore del fenomeno oggetto di indagine, della struttura dell'indagine e dei dati è fondamentale per valutare la convenienza nell'impiego di queste informazioni ausiliarie.

Lo stimatore AR_o, dunque, sembrerebbe preferibile rispetto allo stimatore AC. Tuttavia, i due stimatori proposti non vanno visti in contrapposizione l'uno con l'altro, ma vanno considerati come risposta a due diverse esigenze, una che dà la

priorità al soddisfacimento della condizione di coerenza e l'altra che, libera da questa, ricerca la minimizzazione dell'errore campionario. I risultati ottenuti assumono, quindi, notevole importanza sia da un punto di vista metodologico che da un punto di vista applicativo fornendo delle indicazioni pratiche per un uso consapevole delle informazioni ausiliarie nel processo di stima.

Appendice

A1. Il sistema di pesi dello stimatore CAL

Scriviamo il problema di minimo vincolato nella (1.22) in termini vettoriali

$$\begin{cases} \min_{\mathbf{w}_s} \ (\mathbf{w}_s - \mathbf{d}_s)^t \, \mathbf{\Pi}_s \, (\mathbf{w}_s - \mathbf{d}_s) \\ \mathbf{t}_{\mathbf{X}} - \mathbf{X}_s{}^t \, \mathbf{w}_s \end{cases}$$

La funzione lagrangiana relativa al problema di minimo vincolato è:

$$\mathscr{L}(\mathbf{w}_s, \boldsymbol{\lambda}) = (\mathbf{w}_s - \mathbf{d}_s)^t \, \mathbf{\Pi}_s \, (\mathbf{w}_s - \mathbf{d}_s) - \boldsymbol{\lambda}^t \left(\mathbf{t}_{\mathbf{X}} - \mathbf{X}_s{}^t \, \mathbf{w}_s\right)$$

Le derivate della funzione lagrangiana rispetto $\mathbf{w}_s$ e $\boldsymbol{\lambda}$ sono uguali a

$$\frac{\mathscr{L}(\mathbf{w}_s, \boldsymbol{\lambda})}{\partial \, \mathbf{w}_s} = -2 \, (\mathbf{w}_s - \mathbf{d}_s)^t \, \mathbf{\Pi}_s - \boldsymbol{\lambda}^t \, \mathbf{X}_s{}^t = \mathbf{0} \tag{A.1}$$

$$\frac{\mathscr{L}(\mathbf{w}_s, \boldsymbol{\lambda})}{\partial \, \boldsymbol{\lambda}} = \mathbf{X}_s{}^t \, \mathbf{w}_s - \mathbf{t}_{\mathbf{X}} = \mathbf{0} \tag{A.2}$$

dove $\boldsymbol{\lambda}$ è un vettore di dimension $P \times 1$.

Si moltiplica la (A.1) per $\mathbf{\Pi}_s^{-1} \, \mathbf{X}_s$, quindi si ha

$$-2 \, (\mathbf{w}_s - \mathbf{d}_s)^t \, \mathbf{\Pi}_s \, \mathbf{\Pi}_s^{-1} \, \mathbf{X}_s - \boldsymbol{\lambda}^t \, \mathbf{X}_s{}^t \, \mathbf{\Pi}_s^{-1} \, \mathbf{X}_s = 0$$

e si ricava $\boldsymbol{\lambda}$, ricordando che $\hat{\mathbf{t}}_{\mathbf{X}_{HT}} = \mathbf{X}_s{}^t \, \mathbf{d}_s$ e $\mathbf{t}_{\mathbf{X}} = \mathbf{X}_s{}^t \, \mathbf{w}_s$ per la (A.2),

$$\boldsymbol{\lambda}^t = -2 \, (\mathbf{w}_s - \mathbf{d}_s)^t \, \mathbf{X}_s \, \left(\mathbf{X}_s{}^t \, \mathbf{\Pi}_s^{-1} \, \mathbf{X}_s\right)^{-1}, \tag{A.3}$$

quindi:

$$\begin{aligned} \boldsymbol{\lambda} &= -2 \left(\mathbf{X}_s{}^t \, \mathbf{\Pi}_s^{-1} \, \mathbf{X}_s\right)^{-1} \mathbf{X}_s{}^t \, (\mathbf{w}_s - \mathbf{d}_s) \\ &= -2 \left(\mathbf{X}_s{}^t \, \mathbf{\Pi}_s^{-1} \, \mathbf{X}_s\right)^{-1} \left(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right). \end{aligned}$$

Infine si sostituisce la (A.3) nella (A.1)

$$
\begin{aligned}
&-2\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\boldsymbol{\Pi}_s+2\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t=0\\
&-2\,\mathbf{w}_s{}^t\,\boldsymbol{\Pi}_s+2\,\mathbf{d}_s{}^t\,\boldsymbol{\Pi}_s+2\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\right)^{-1}\mathbf{X}_s^t=0\\
&\mathbf{w}_s{}^t\,\boldsymbol{\Pi}_s=\mathbf{d}_s{}^t\,\boldsymbol{\Pi}_s+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\right)^{-1}\mathbf{X}_s^t\\
&\mathbf{w}_s{}^t=\mathbf{d}_s^t+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}
\end{aligned}
$$

Quindi l'espressione del vettore dei pesi dello stimatore CAL è

$$
\mathbf{w}_{CAL}=\mathbf{d}_s+\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\left(\mathbf{X}_s{}^t\,\boldsymbol{\Pi}_s^{-1}\,\mathbf{X}_s\right)^{-1}\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right).
$$

A2. Il sistema di pesi dello stimatore $GREG_o$ di Montanari (1987)

Il sistema di pesi dello stimatore $GREG_o$ proposto da Montanari (1987) può essere ricavato risolvendo un problema di minimo vincolato in cui, diversamente da quanto visto in Appendice A1, la metrica con cui è misurata la distanza dei pesi è $\boldsymbol{\Delta}_s^{-1}$. Il problema di minimo vincolato illustrato anche nella (1.25) in termini vettoriali

$$
\begin{cases}
\min_{\mathbf{w}_s}\ \left(\mathbf{w}_s-\mathbf{d}_s\right)^t\underline{\boldsymbol{\Delta}}_s^{-1}\left(\mathbf{w}_s-\mathbf{d}_s\right)\\
\mathbf{t}_{\mathbf{X}}-\mathbf{X}_s^t\,\mathbf{w}_s
\end{cases}
$$

La funzione lagrangiana relativa al problema di minimo vincolato è

$$
\mathscr{L}\left(\mathbf{w}_s,\boldsymbol{\lambda}\right)=\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\underline{\boldsymbol{\Delta}}_s^{-1}\left(\mathbf{w}_s-\mathbf{d}_s\right)-\boldsymbol{\lambda}^t\left(\mathbf{t}_{\mathbf{X}}-\mathbf{X}_s{}^t\,\mathbf{w}_s\right)
$$

Le derivate della funzione lagrangiana rispetto $\mathbf{w}_s$ e $\boldsymbol{\lambda}$ sono uguali a

$$
\frac{\mathscr{L}\left(\mathbf{w}_s,\boldsymbol{\lambda}\right)}{\partial\,\mathbf{w}_s}=-2\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\underline{\boldsymbol{\Delta}}_s^{-1}-\boldsymbol{\lambda}^t\,\mathbf{X}_s{}^t=\mathbf{0} \tag{A.4}
$$

$$
\frac{\mathscr{L}\left(\mathbf{w}_s,\boldsymbol{\lambda}\right)}{\partial\,\boldsymbol{\lambda}}=\mathbf{X}_s{}^t\,\mathbf{w}_s-\mathbf{t}_{\mathbf{X}}=\mathbf{0} \tag{A.5}
$$

dove $\boldsymbol{\lambda}$ è un vettore di dimension $P\times 1$.

Si moltiplica la (A.4) per $\underline{\boldsymbol{\Delta}}_s\,\mathbf{X}_s$, quindi si ha

$$
-2\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\underline{\boldsymbol{\Delta}}_s^{-1}\,\underline{\boldsymbol{\Delta}}_s\,\mathbf{X}_s-\boldsymbol{\lambda}^t\,\mathbf{X}_s{}^t\,\underline{\boldsymbol{\Delta}}_s\,\mathbf{X}_s=0
$$

e si ricava $\boldsymbol{\lambda}$, ricordando che $\hat{\mathbf{t}}_{\mathbf{X}_{HT}}=\mathbf{X}_s{}^t\,\mathbf{d}_s$ e $\mathbf{t}_{\mathbf{X}}=\mathbf{X}_s{}^t\,\mathbf{w}_s$ per la (A.5),

$$
\boldsymbol{\lambda}^t=-2\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\mathbf{X}_s\,\left(\mathbf{X}_s{}^t\,\underline{\boldsymbol{\Delta}}_s\,\mathbf{X}_s\right)^{-1}, \tag{A.6}
$$

quindi:

$$\begin{aligned}\boldsymbol{\lambda} &= -2\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t\left(\mathbf{w}_s-\mathbf{d}_s\right)\\ &= -2\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right).\end{aligned}$$

Infine si sostituisce la (A.6) nella (A.4):

$$\begin{aligned}&-2\left(\mathbf{w}_s-\mathbf{d}_s\right)^t\underline{\boldsymbol{\Delta}}_s^{-1}+2\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t=0\\ &-2\ \mathbf{w}_s{}^t\ \underline{\boldsymbol{\Delta}}_s^{-1}+2\ \mathbf{d}_s{}^t\ \underline{\boldsymbol{\Delta}}_s^{-1}+2\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t=0\\ &\mathbf{w}_s{}^t\ \underline{\boldsymbol{\Delta}}_s^{-1}=\mathbf{d}_s{}^t\ \underline{\boldsymbol{\Delta}}_s^{-1}+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t\\ &\mathbf{w}_s{}^t=\mathbf{d}_s{}^t+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\end{aligned}$$

Quindi l'espressione del vettore dei dei pesi dello stimatore $GREG_o$ è

$$\mathbf{w}_{GREG_o}=\mathbf{d}_s+\underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\left(\mathbf{X}_s{}^t\ \underline{\boldsymbol{\Delta}}_s\ \mathbf{X}_s\right)^{-1}\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right).$$

A3. Il sistema di pesi dello stimatore AC

Scriviamo il problema di minimo vincolato nella (2.1) in termini vettoriali

$$\begin{cases}\min_{\mathbf{w}_2}\ \left(\mathbf{w}_2-\mathbf{d}_2\right)^t\boldsymbol{\Pi}_2\left(\mathbf{w}_2-\mathbf{d}_2\right)\\ \begin{pmatrix}\mathbf{t}_{\mathbf{X}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}\end{pmatrix}-\begin{pmatrix}\mathbf{X}_2{}^t\ \mathbf{w}_2\\ \mathbf{Z}_2{}^t\ \mathbf{w}_2\end{pmatrix}\end{cases}$$

La funzione lagrangiana relativa al problema di minimo vincolato è:

$$\mathscr{L}\left(\mathbf{w}_2,\boldsymbol{\lambda}\right)=\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\boldsymbol{\Pi}_2\left(\mathbf{w}_2-\mathbf{d}_2\right)-\boldsymbol{\lambda}^t\begin{pmatrix}\mathbf{t}_{\mathbf{X}}-\mathbf{X}_2{}^t\ \mathbf{w}_2\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\mathbf{Z}_2{}^t\ \mathbf{w}_2\end{pmatrix}$$

Le derivate della funzione lagrangiana rispetto $\mathbf{w}_2$ e $\boldsymbol{\lambda}$ sono uguali a:

$$\frac{\mathscr{L}\left(\mathbf{w}_2,\boldsymbol{\lambda}\right)}{\partial\ \mathbf{w}_2}=-2\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\boldsymbol{\Pi}_2-\boldsymbol{\lambda}^t\begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}=\mathbf{0}\tag{A.7}$$

$$\frac{\mathscr{L}\left(\mathbf{w}_2,\boldsymbol{\lambda}\right)}{\partial\ \boldsymbol{\lambda}}=\begin{pmatrix}\mathbf{t}_{\mathbf{X}}-\mathbf{X}_2{}^t\ \mathbf{w}_2\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\mathbf{Z}_2{}^t\ \mathbf{w}_2\end{pmatrix}=\mathbf{0}\tag{A.8}$$

dove $\boldsymbol{\lambda}$ è un vettore di dimension $(P+M)\times 1$.

Si moltiplica la (A.7) per $\boldsymbol{\Pi}_2^{-1}\left(\mathbf{X}_2\ \mathbf{Z}_2\right)$, quindi si ha

$$-2\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\boldsymbol{\Pi}_2\ \boldsymbol{\Pi}_2^{-1}\ \left(\mathbf{X}_2\ \mathbf{Z}_2\right)-\boldsymbol{\lambda}^t\ \begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}\boldsymbol{\Pi}_2^{-1}\left(\mathbf{X}_2\ \mathbf{Z}_2\right)=\mathbf{0}$$

e da questa si ricava $\boldsymbol{\lambda}$. Indicando con

$$\hat{\mathbf{t}}_{\mathbf{X}_{HT}} = \mathbf{X}_2{}^t\ \mathbf{d}_2$$

$$\hat{\mathbf{t}}_{\mathbf{Z}_{HT}} = \mathbf{Z}_2{}^t\ \mathbf{d}_2,$$

$$\mathbf{t}_{\mathbf{X}} = \mathbf{X}_2{}^t\ \mathbf{w}_{AC},$$

$$\widetilde{\mathbf{t}}_{\mathbf{Z}_1} = \mathbf{Z}_2{}^t\ \mathbf{w}_{AC}$$

e definendo

$$\mathbf{Q} = \begin{bmatrix} \mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2 & \mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{Z}_2 \\ \mathbf{Z}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2 & \mathbf{Z}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{Z}_2 \end{bmatrix} \tag{A.9}$$

si ha che

$$\boldsymbol{\lambda}^t = -2\,(\mathbf{w}_{AC} - \mathbf{d}_2)^t\,(\mathbf{X}_2\ \mathbf{Z}_2)\ \mathbf{Q}^{-1}, \tag{A.10}$$

quindi:

$$\begin{aligned} \boldsymbol{\lambda} &= -2\ \mathbf{Q}^{-1} \begin{pmatrix} \mathbf{X}_2{}^t \\ \mathbf{Z}_2{}^t \end{pmatrix} (\mathbf{w}_2 - \mathbf{d}_2) \\ &= -2\ \mathbf{Q}^{-1} \begin{pmatrix} \mathbf{t}_{\mathbf{X}} - \mathbf{X}_2{}^t\ \mathbf{w}_2 \\ \widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{Z}_2{}^t\ \mathbf{w}_2 \end{pmatrix}. \end{aligned}$$

L'inversa della matrice a blocchi $\mathbf{Q}$ definita nella (A.9) è

$$\mathbf{Q}^{-1} = \begin{bmatrix} \left(\mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2\right)^{-1} + \hat{\mathbf{L}}\hat{\mathbf{M}}\hat{\mathbf{L}}^t & -\hat{\mathbf{L}}\hat{\mathbf{M}} \\ \hat{\mathbf{M}}\hat{\mathbf{L}}^t & \hat{\mathbf{M}} \end{bmatrix}$$

con

$$\hat{\mathbf{L}} = \left(\mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2\right)^{-1} \mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{Z}_2 \tag{A.11}$$

$$\hat{\mathbf{M}} = \left(\mathbf{Z}_2{}^t\ \hat{\mathbf{R}}\ \mathbf{Z}_2\right)^{-1} \tag{A.12}$$

dove

$$\hat{\mathbf{R}} = \mathbf{\Pi}_2^{-1}\left(\mathbf{I} - \hat{\mathbf{P}}\right)$$

in cui

$$\hat{\mathbf{P}} = \mathbf{X}_2\left(\mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}\ \mathbf{X}_2\right)^{-1} \mathbf{X}_2{}^t\ \mathbf{\Pi}_2^{-1}$$

è la matrice degli operatori di proiezione sullo spazio generato dalle colonne della matrice $\mathbf{X}_2$.

Se si sostituisce la (A.10) nella (A.7)

$$
\begin{aligned}
&-2\,(\mathbf{w}_2-\mathbf{d}_2)^t\,\mathbf{\Pi}_2+2\begin{pmatrix}\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\end{pmatrix}^t\mathbf{Q}^{-1}\begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}=\mathbf{0}\\
&-2\;\mathbf{w}_2{}^t\;\mathbf{\Pi}_2+2\;\mathbf{d}_2{}^t\;\mathbf{\Pi}_2+2\begin{pmatrix}\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\end{pmatrix}^t\mathbf{Q}^{-1}\begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}=\mathbf{0}\\
&\mathbf{w}_2{}^t\;\mathbf{\Pi}_2=\;\mathbf{d}_2{}^t\;\mathbf{\Pi}_2+\begin{pmatrix}\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\end{pmatrix}^t\mathbf{Q}^{-1}\begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}\\
&\mathbf{w}_2{}^t=\mathbf{d}_2{}^t+\begin{pmatrix}\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\end{pmatrix}^t\mathbf{Q}^{-1}\begin{pmatrix}\mathbf{X}_2{}^t\\ \mathbf{Z}_2{}^t\end{pmatrix}\mathbf{\Pi}_2{}^t\\
&\mathbf{w}_2=\mathbf{d}_2+\mathbf{\Pi}_2^{-1}\,(\mathbf{X}_2\;\mathbf{Z}_2)\,\mathbf{Q}^{-1}\begin{pmatrix}\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\\ \tilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\end{pmatrix}
\end{aligned}
\tag{A.13}
$$

La (A.13) può essere semplificata sviluppando il prodotto $\mathbf{\Pi}_2^{-1}\,(\mathbf{X}_2\;\mathbf{Z}_2)\,\mathbf{Q}^{-1}$. Infatti,

$$
\begin{aligned}
\mathbf{w}_2 \;=\;& \mathbf{d}_2+\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\left(\mathbf{X}_2{}^t\;\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\right)^{-1}\left(\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\\
&+\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\;\hat{\mathbf{L}}\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\\
&-\mathbf{\Pi}_2^{-1}\;\mathbf{Z}_2\;\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t_X}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\\
&-\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\;\mathbf{L}\mathbf{M}\left(\mathbf{t_Z}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)\\
&+\mathbf{\Pi}_2^{-1}\;\mathbf{Z}_2\;\hat{\mathbf{M}}\left(\mathbf{t_Z}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right).
\end{aligned}
\tag{A.14}
$$

Considerando che

$$
\begin{aligned}
\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\;\hat{\mathbf{L}} \;&=\; \mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\left(\mathbf{X}_2{}^t\;\mathbf{\Pi}_2^{-1}\;\mathbf{X}_2\right)^{-1}\mathbf{X}_2{}^t\;\mathbf{\Pi}_2^{-1}\;\mathbf{Z}_2\\
&=\; \hat{\mathbf{P}}\;\mathbf{\Pi}_2^{-1}\;\mathbf{Z}_2,
\end{aligned}
$$

la (A.14) diventa

$$\begin{aligned}
\mathbf{w}_2 &= \mathbf{d}_2 + \mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad + \hat{\mathbf{P}}\,\mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2{}^t\,\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad - \mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad - \hat{\mathbf{P}}\,\mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2{}^t\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\
&\quad + \mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\
&= \mathbf{d}_2 + \mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad - \left(\mathbf{I} - \hat{\mathbf{P}}\right)\mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad + \left(\mathbf{I} - \hat{\mathbf{P}}\right)\mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\
&= \mathbf{d}_2 + \mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad - \hat{\mathbf{R}}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\hat{\mathbf{L}}^t\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad + \hat{\mathbf{R}}\,\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \\
&= \mathbf{d}_2 + \mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) \\
&\quad - \mathbf{R}\,\mathbf{Z}_2\,\mathbf{M}\left(\mathbf{L}^t\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) - \mathbf{t}_\mathbf{Z} + \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right) \qquad \text{(A.15)}
\end{aligned}$$

Ricordando che

$$\hat{\mathbf{t}}_{\mathbf{Z}_{HT}} = \mathbf{Z}_2{}^t\,\mathbf{d}_2$$

$$\mathbf{w}_{CAL} = \mathbf{d}_2 + \mathbf{\Pi}_2\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)$$

e sostituendo la (A.11) e la (A.12) nella (A.15) si ha

$$\begin{aligned}
\mathbf{w}_2 &= \mathbf{w}_{CAL} - \hat{\mathbf{R}}\;\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{Z}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right) - \mathbf{t}_\mathbf{Z} + \mathbf{Z}_2{}^t\,\mathbf{d}_2\right) \\
&= \mathbf{w}_{CAL} + \hat{\mathbf{R}}\;\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \mathbf{Z}_2{}^t\left(\mathbf{d}_2 + \mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\left(\mathbf{X}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{X}_2\right)^{-1}\left(\mathbf{t}_\mathbf{X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)\right)\right) \\
&= \mathbf{w}_{CAL} + \hat{\mathbf{R}}\;\mathbf{Z}_2\,\hat{\mathbf{M}}\left(\mathbf{t}_\mathbf{Z} - \mathbf{Z}_2{}^t\,\mathbf{w}_{CAL}\right)
\end{aligned}$$

Quindi l'espressione del vettore dei dei pesi dello stimatore AC è

$$\mathbf{w}_{AC} = \mathbf{w}_{CAL} + \hat{\mathbf{R}}\;\mathbf{Z}_2\left(\mathbf{Z}_2{}^t\,\mathbf{\Pi}_2^{-1}\,\mathbf{Z}_2\right)^{-1}\left(\mathbf{t}_\mathbf{Z} - \mathbf{Z}_2{}^t\,\mathbf{w}_{CAL}\right).$$

A4. La covarianza A_1, A_2

La covarianza di A_1, A_2 è uguale a $\mathbb{E}\left[A_1\ A_2\right]$ poichè $\mathbb{E}[A_2] = \mathbf{0}$

$$\begin{aligned}
&\mathbb{E}\left[A_1\ A_2\right] \\
&= \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \left(\mathbf{t_U} - \check{\mathbf{t}}_\mathbf{U}\right)^t \boldsymbol{\beta}\right] \\
&= \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \mathbf{t_U}^t\boldsymbol{\beta} - \hat{t}_{Y_{GREG}}\ \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right] \\
&= \mathbb{E}\left[\hat{t}_{Y_{GREG}}\right]\mathbf{t_U}^t\boldsymbol{\beta} - \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right] \\
&= \hat{t}_Y\ \mathbf{t_U}^t\boldsymbol{\beta} - \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right]. \qquad \text{(A.16)}
\end{aligned}$$

Continuando a sviluppare l'espressione si ha, invece,

$$\begin{aligned}
&\mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \mathbf{t_U}^t\boldsymbol{\beta} - \hat{t}_{Y_{GREG}}\ \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right] \\
&= \mathbb{E}\left[\left(\hat{t}_{Y_{HT}} + \mathbf{t_U}^t\boldsymbol{\beta} - \hat{\mathbf{t}}_{\mathbf{U}_{HT}}{}^t\boldsymbol{\beta}\right)^t\left(\mathbf{t_U}^t\boldsymbol{\beta} - \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right)\right] \\
&= \mathbb{E}\left[\hat{t}_{Y_{HT}}\mathbf{t_U}^t\boldsymbol{\beta} - \hat{t}_{Y_{HT}}\check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t\mathbf{t_U}\mathbf{t_U}^t\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{t_U}\check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta} - \boldsymbol{\beta}^t\hat{\mathbf{t}}_{\mathbf{U}_{HT}}\mathbf{t_U}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t\hat{\mathbf{t}}_{\mathbf{U}_{HT}}\check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right] \\
&= \hat{t}_Y\mathbf{t_U}^t\boldsymbol{\beta} - \mathbb{E}\left[\hat{t}_{Y_{HT}}\check{\mathbf{t}}_\mathbf{U}{}^t\right]\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{t_U}\mathbf{t_U}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t\ \mathbb{E}\left[\hat{\mathbf{t}}_{\mathbf{U}_{HT}}\check{\mathbf{t}}_\mathbf{U}{}^t\right]\boldsymbol{\beta}. \qquad \text{(A.17)}
\end{aligned}$$

Poiché

$$Cov\left(\hat{t}_{Y_{GREG}}, \check{\mathbf{t}}_\mathbf{U}\right) = \mathbb{E}\left[\hat{t}_{Y_{GREG}}\ \check{\mathbf{t}}_\mathbf{U}{}^t\boldsymbol{\beta}\right] - \hat{t}_Y\ \mathbf{t_U}^t\boldsymbol{\beta}$$

e

$$Cov\left(\hat{t}_{Y_{HT}}, \check{\mathbf{t}}_\mathbf{U}\right) = \mathbb{E}\left[\hat{t}_{Y_{HT}}\ \check{\mathbf{t}}_\mathbf{U}\right] - t_Y\ \mathbf{t_U}^t$$

$$Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \check{\mathbf{t}}_\mathbf{U}\right) = \mathbb{E}\left[\hat{\mathbf{t}}_{\mathbf{U}_{HT}}\ \check{\mathbf{t}}_\mathbf{U}\right] - \mathbf{t_U}\ \mathbf{t_U}^t$$

Dalla (A.16) si ha che

$$Cov\left(A_1, A_2\right) = -Cov\left(\hat{t}_{Y_{GREG}}, \check{\mathbf{t}}_\mathbf{U}\right)^t\boldsymbol{\beta} \qquad \text{(A.18)}$$

e, in maniera equivalente, dalla (A.17)

$$Cov\left(A_1, A_2\right) = \boldsymbol{\beta}^t\ Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \check{\mathbf{t}}_\mathbf{U}\right)\boldsymbol{\beta} - Cov\left(\hat{t}_{Y_{HT}}, \check{\mathbf{t}}_\mathbf{U}\right)^t\boldsymbol{\beta}. \qquad \text{(A.19)}$$

Poiché in $\check{\mathbf{t}}_\mathbf{U}{}^t = \left(\mathbf{t_X}^t\ \mathbf{t_{Z_1}}{}^t\right)\ \mathbf{t_X}$ è una costante rispetto al campione, la (A.18) può essere semplificata in quanto

$$\begin{aligned}
Cov\left(\hat{t}_{Y_{GREG}}, \check{\mathbf{t}}_\mathbf{U}\right)\boldsymbol{\beta} &= \left[\ Cov\left(\hat{t}_{Y_{GREG}}, \mathbf{t_X}\right)\ \ Cov\left(\hat{t}_{Y_{GREG}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\ \right]\begin{bmatrix}\boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M\end{bmatrix} \\
&= \left[\ \mathbf{0}\ \ Cov\left(\hat{t}_{Y_{GREG}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\ \right]\begin{bmatrix}\boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M\end{bmatrix} \\
&= Cov\left(\hat{t}_{Y_{GREG}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M
\end{aligned}$$

Per lo stesso motivo la (A.19) può essere semplificata in quanto

$$\begin{aligned} Cov\left(\hat{t}_{Y_{HT}}, \check{\mathbf{t}}_{\mathbf{U}}\right)\boldsymbol{\beta} &= \left[\begin{array}{cc} Cov\left(\hat{t}_{Y_{HT}}, \mathbf{t}_{\mathbf{X}}\right) & Cov\left(\hat{t}_{Y_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M \end{array}\right] \\ &= \left[\begin{array}{cc} \mathbf{0} & Cov\left(\hat{t}_{Y_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M \end{array}\right] \\ &= Cov\left(\hat{t}_{Y_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M \end{aligned}$$

$$\begin{aligned} \boldsymbol{\beta}^t\, Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \check{\mathbf{t}}_{\mathbf{X}}\right)\boldsymbol{\beta} &= \left[{\boldsymbol{\beta}_P}^t\ {\boldsymbol{\beta}_M}^t\right] \left[\begin{array}{cc} Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \mathbf{t}_{\mathbf{X}}\right) & Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \\ Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \mathbf{t}_{\mathbf{X}}\right) & Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M \end{array}\right] \\ &= \left[{\boldsymbol{\beta}_P}^t\ {\boldsymbol{\beta}_M}^t\right] \left[\begin{array}{cc} \mathbf{0} & Cov\left(\hat{\mathbf{t}}_{\mathbf{X}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \\ \mathbf{0} & Cov\left(\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_P \\ \boldsymbol{\beta}_M \end{array}\right] \\ &= \boldsymbol{\beta}^t\, Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M \end{aligned}$$

Quindi, in maniera equivalente, abbiamo che

$$\begin{aligned} Cov\left(A_1, A_2\right) &= -Cov\left(\hat{t}_{Y_{GREG}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M \\ &= \boldsymbol{\beta}^t\, Cov\left(\hat{\mathbf{t}}_{\mathbf{U}_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M - Cov\left(\hat{t}_{Y_{HT}}, \tilde{\mathbf{t}}_{\mathbf{Z}_1}\right)\boldsymbol{\beta}_M. \end{aligned}$$

A5. Il vettore $\boldsymbol{\beta}$ che minimizza la varianza da disegno dello stimatore AR con solo variabili ausiliarie campionarie

Lo stimatore AR con sole variabili ausiliarie campionarie, presentato nella (3.3) è

$$\hat{t}_{Y_{AR}} = \hat{t}_{Y_{HT}} + \left(\tilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \boldsymbol{\gamma}. \tag{A.20}$$

la sua varianza da disegno è

$$\begin{aligned} Var\left(\hat{t}_{Y_{AR}}\right) &= \mathbf{y}^t\, \boldsymbol{\Delta}_2\, \mathbf{y} \\ &\quad + \boldsymbol{\gamma}^t\mathbf{Z}^t\, \boldsymbol{\Delta}_1\, \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\gamma}^t\mathbf{Z}^t\, \boldsymbol{\Delta}_2\, \mathbf{Z}\boldsymbol{\gamma} \\ &\quad + 2\, \mathbf{y}^t\, \boldsymbol{\Delta}_{21}\, \mathbf{Z}\boldsymbol{\gamma} - 2\, \mathbf{y}^t\, \boldsymbol{\Delta}_2\, \mathbf{Z}\boldsymbol{\gamma} \\ &\quad - 2\, \boldsymbol{\gamma}^t\mathbf{Z}^t\, \boldsymbol{\Delta}_{12}\, \mathbf{Z}\boldsymbol{\gamma}. \end{aligned} \tag{A.21}$$

Il valore del regressore $\boldsymbol{\gamma}$, considerato come una costante, che minimizza la varianza di (A.20) è ottenuto derivando la (A.21) e uguagliando la derivata a 0, quindi

$$\begin{aligned}\frac{\partial\, Var\left(\hat{t}_{Y_{AR}}\right)}{\partial\, \boldsymbol{\gamma}} &= \boldsymbol{\gamma}^t \mathbf{Z}^t\, \boldsymbol{\Delta}_1\, \mathbf{Z} + \boldsymbol{\gamma}^t \mathbf{Z}^t\, \boldsymbol{\Delta}_2\, \mathbf{Z} - 2\, \boldsymbol{\gamma} \mathbf{Z}^t\, \boldsymbol{\Delta}_{12}\, \mathbf{Z} \\ &\quad + 2\, \mathbf{y}^t\, \boldsymbol{\Delta}_{21}\, \mathbf{Z} - 2\, \mathbf{y}^t\, \boldsymbol{\Delta}_2\, \mathbf{Z} = \mathbf{0}\end{aligned}$$

e

$$\boldsymbol{\gamma}_o = \left(\mathbf{Z}^t\, \boldsymbol{\Delta}_1\, \mathbf{Z} + \mathbf{Z}^t\, \boldsymbol{\Delta}_2\, \mathbf{Z} - 2\, \mathbf{Z}^t\, \boldsymbol{\Delta}_{12}\, \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\, \boldsymbol{\Delta}_{12}\, \mathbf{y} - \mathbf{Z}^t\, \boldsymbol{\Delta}_2\, \mathbf{y}\right).$$

A6. Il sistema di pesi dello stimatore AR_o con sole variabili ausiliarie campionarie

Il sistema di pesi dello stimatore AR_{o_z} nella (3.5) può essere ricavato risolvendo un problema di minimo vincolato simile a quello visto in Appendice A2, in cui, però, il vincolo non è $\mathbf{t_X} - \mathbf{X_2}^t\, \mathbf{w}_2$ ma

$$\hat{\boldsymbol{\Gamma}}\, \widetilde{\mathbf{t}}_{\mathbf{Z}_1} + \left(\mathbf{I} - \hat{\boldsymbol{\Gamma}}\right) \mathbf{Z_2}^t\, \mathbf{d}_2 - \mathbf{Z_2}^t\, \mathbf{w}_2$$

Da questa espressione abbiamo che

$$(\mathbf{d}_2 - \mathbf{w}_2)^t\, \mathbf{Z}_2 = \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \mathbf{d_2}^t\, \mathbf{Z}_2\right) \hat{\boldsymbol{\Gamma}} \tag{A.22}$$

Quindi

$$\begin{cases} \min_{\mathbf{w}_2}\ (\mathbf{w}_2 - \mathbf{d}_2)^t\, \underline{\boldsymbol{\Delta}}_2^{-1}\, (\mathbf{w}_2 - \mathbf{d}_2) \\ \hat{\boldsymbol{\Gamma}}\, \widetilde{\mathbf{t}}_{\mathbf{Z}_1} + \left(\mathbf{I} - \hat{\boldsymbol{\Gamma}}\right) \mathbf{Z_2}^t\, \mathbf{d}_2 - \mathbf{Z}_s{}^t\, \mathbf{w}_2 = \mathbf{0} \end{cases}$$

La funzione lagrangiana relativa al problema di minimo vincolato è

$$\mathscr{L}\,(\mathbf{w}_2, \boldsymbol{\lambda}) = (\mathbf{w}_2 - \mathbf{d}_2)^t\, \underline{\boldsymbol{\Delta}}_2^{-1}\, (\mathbf{w}_2 - \mathbf{d}_2) - \boldsymbol{\lambda}^t \left(\hat{\boldsymbol{\Gamma}}\, \widetilde{\mathbf{t}}_{\mathbf{Z}_1} + \left(\mathbf{I} - \hat{\boldsymbol{\Gamma}}\right) \mathbf{Z_2}^t\, \mathbf{d}_2 - \mathbf{Z}_s{}^t\, \mathbf{w}_2\right)$$

Le derivate della funzione lagrangiana rispetto $\mathbf{w}_2$ e $\boldsymbol{\lambda}$ sono uguali a:

$$\frac{\mathscr{L}\,(\mathbf{w}_2, \boldsymbol{\lambda})}{\partial\, \mathbf{w}_2} = -2\, (\mathbf{w}_2 - \mathbf{d}_2)^t\, \underline{\boldsymbol{\Delta}}_2^{-1} - \boldsymbol{\lambda}^t\, \mathbf{Z_2}^t = \mathbf{0} \tag{A.23}$$

$$\frac{\mathscr{L}\,(\mathbf{w}_s, \boldsymbol{\lambda})}{\partial\, \boldsymbol{\lambda}} = \hat{\boldsymbol{\Gamma}}\, \widetilde{\mathbf{t}}_{\mathbf{Z}_1} + \left(\mathbf{I} - \hat{\boldsymbol{\Gamma}}\right) \mathbf{Z_2}^t\, \mathbf{d}_2 - \mathbf{Z}_s{}^t\, \mathbf{w}_2 = \mathbf{0} \tag{A.24}$$

dove $\boldsymbol{\lambda}$ è un vettore di dimension $M \times 1$.

Si moltiplica la (A.23) per $\underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2$, quindi si ha

$$-2\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\ \underline{\boldsymbol{\Delta}}_2^{-1}\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2-\boldsymbol{\lambda}^t\ {\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2=0$$

e si ricava $\boldsymbol{\lambda}$. Ricordando, oltre al risultato nella (A.22), che $\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}={\mathbf{Z}_2}^t\ \mathbf{d}_2$ si ha

$$\begin{aligned}\boldsymbol{\lambda}^t &= -2\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\mathbf{Z}_s\ \left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}\\ &= -2\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)\hat{\boldsymbol{\Gamma}}\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}\end{aligned}\qquad(A.25)$$

quindi:

$$\boldsymbol{\lambda} = -2\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}\hat{\boldsymbol{\Gamma}}\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)$$

Infine sostituendo la (A.25) nella (A.23):

$$\begin{aligned}&-2\left(\mathbf{w}_2-\mathbf{d}_2\right)^t\ {\underline{\boldsymbol{\Delta}}_2}^{-1}+2\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)\hat{\boldsymbol{\Gamma}}\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}{\mathbf{Z}_2}^t=0\\ &-2\ {\mathbf{w}_2}^t\ \underline{\boldsymbol{\Delta}}_2^{-1}+2\ {\mathbf{d}_2}^t\ \underline{\boldsymbol{\Delta}}_2^{-1}+2\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)\hat{\boldsymbol{\Gamma}}\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}{\mathbf{Z}_2}^t=0\\ &{\mathbf{w}_2}^t\ \underline{\boldsymbol{\Delta}}_2^{-1}={\mathbf{d}_2}^t\ \underline{\boldsymbol{\Delta}}_2^{-1}+\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)\hat{\boldsymbol{\Gamma}}\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}{\mathbf{Z}_2}^t\\ &{\mathbf{w}_2}^t={\mathbf{d}_2}^t+\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right)\hat{\boldsymbol{\Gamma}}\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}{\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2.\end{aligned}$$

Quindi l'espressione del vettore dei pesi dello stimatore AR_{o_x} con le sole variabili ausiliarie non-campionarie è

$$\mathbf{w}_{AR_{o_x}}=\mathbf{d}_2+\underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\left({\mathbf{Z}_2}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2\right)^{-1}\hat{\boldsymbol{\Gamma}}\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-{\mathbf{d}_2}^t\ \mathbf{Z}_2\right).$$

A7. I vettori $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ che minimizzano la varianza da disegno dello stimatore AR con informazioni ausiliarie non-campionarie e campionarie

Lo stimatore AR che considera sia informazioni ausiliarie non-campionarie e campionarie, già presentato nella (3.1), è

$$\hat{t}_{Y_{AR}}=\hat{t}_{Y_{HT}}+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\hat{\boldsymbol{\beta}}+\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t\hat{\boldsymbol{\gamma}}.$$

Consideriamo, però una sua espressione equivalente, anche questa già presentata nella (3.11),

$$\hat{t}_{Y_{AR}}=\hat{t}_{Y_{HT}}+\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\hat{\boldsymbol{\beta}}-\left(\mathbf{t}_{\mathbf{X}}-\hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t\hat{\mathbf{L}}\hat{\boldsymbol{\gamma}}+\left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1}-\hat{\mathbf{t}}_{\mathbf{Z}_{2_{HT}}}\right)^t\hat{\boldsymbol{\gamma}},\quad(A.26)$$

in cui $\hat{\mathbf{L}} = (\mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{X}_2)^{-1}\,\mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2$ è la stima della matrice dei coefficienti di regressione del set di variabili Z sullo spazio generato dalle colonne della matrice $\mathbf{X}_2$.

La varianza da disegno dell'espressione "ortogonalizzata" dello stimatore AR espresso nella (3.13)

$$\begin{aligned} Var\left(\hat{t}_{Y_{AR}}\right) \;=\; & \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} \\ & + \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \ \boldsymbol{\gamma}^t\mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L}\boldsymbol{\gamma} - \ 2\ \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} + \ 2\ \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & + 2\ \boldsymbol{\gamma}^t\mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \boldsymbol{\gamma}^t\mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}. \end{aligned} \tag{A.27}$$

Dalle equazioni normali

i) $\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \left[\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} - \mathbf{X}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right] = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$

ii) $\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\boldsymbol{\gamma} + \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y} + \mathbf{Z}^t\ \boldsymbol{\Delta}_2\mathbf{P}\ \mathbf{y} = \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$

Poichè

- $\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$
- $\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}$

si ha che

- $\mathbf{X}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \mathbf{0}$.

La (A.27) si riduce a

$$\begin{aligned} Var\left(\hat{t}_{Y_{AR}}\right) \;=\; & \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} \\ & + \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - \boldsymbol{\gamma}^t\mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\boldsymbol{\gamma}. \end{aligned} \tag{A.28}$$

L'espressione di $\boldsymbol{\beta}$, considerato come una costante, che minimizza la varianza dello stimatore nella (3.1) è ottenuta derivando la (A.28) e uguagliando la derivata a 0

$$\frac{\partial\ Var\left(\hat{t}_{Y_{AR}}\right)}{\partial\ \boldsymbol{\beta}} = 2\ \boldsymbol{\beta}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X} - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}$$

quindi

$$\boldsymbol{\beta}_o = \left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$$

che è esattamente la quantità che minimizza la varianza dello stimatore $GREG_o$. La derivata della (A.27) rispetto a $\boldsymbol{\gamma}$ è

$$\begin{aligned}\frac{\partial\ Var\left(\hat{t}_{Y_{AR}}\right)}{\partial\ \boldsymbol{\gamma}} &= 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} \\ &\quad + 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z} \\ &\quad - 2\ \mathbf{L}\mathbf{X}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} \\ &\quad - 4\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \mathbf{0}\end{aligned}$$

e l'espressione di $\boldsymbol{\gamma}$, sempre considerato come una costante, che minimizza la varianza da disegno è

$$\begin{aligned}\boldsymbol{\gamma}_o &= \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \\ &\quad \times \left(\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right) \\ &= \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z} - 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{y}\right)\end{aligned}$$

poiché

$$\begin{aligned}\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} - \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} \\ &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\right)\mathbf{Z} \\ &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Z} \\ &= \mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}\end{aligned}$$

e in maniera equivalente

$$\begin{aligned}\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{L}^t\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} \\ &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2\right)\mathbf{y} \\ &= \mathbf{Z}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I} - \mathbf{P}\right)\mathbf{y} \\ &= \mathbf{Z}^t\ \mathbf{R}\ \mathbf{y}\end{aligned}$$

in quanto $\mathbf{R} = \boldsymbol{\Delta}_2\,(\mathbf{I}$ - $\mathbf{P})$, con $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1}\mathbf{X}^t\ \boldsymbol{\Delta}_2$.

A8. La varianza da disegno dello stimatore AR con informazioni ausiliarie non-campionarie e campionarie

Lo stimatore AR_o che considera sia informazioni ausiliarie non-campionarie e campionarie, già presentato nella (3.2), è

$$\hat{t}_{Y_{AR_o}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}_o + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}_o.$$

Consideriamo, come fatto in appendice A7, una sua espressione equivalente

$$\hat{t}_{Y_{AR_o}} = \hat{t}_{Y_{HT}} + \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\boldsymbol{\beta}}_o - \left(\mathbf{t_X} - \hat{\mathbf{t}}_{\mathbf{X}_{HT}}\right)^t \hat{\mathbf{L}}\hat{\boldsymbol{\gamma}}_o + \left(\widetilde{\mathbf{t}}_{\mathbf{Z}_1} - \hat{\mathbf{t}}_{\mathbf{Z}_{HT}}\right)^t \hat{\boldsymbol{\gamma}}_o,$$

in cui $\hat{\mathbf{L}} = \left(\mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{X}_2\right)^{-1} \mathbf{X}_2{}^t\ \underline{\boldsymbol{\Delta}}_2\ \mathbf{Z}_2$ è la stima della matrice dei coefficienti di regressione del set di variabili Z sullo spazio generato dalle colonne della matrice $\mathbf{X}_2$.

La varianza da disegno dello stimatore AR_o, già presentata in Appendice A7, che riportiamo qui per comodità è uguale a

$$\begin{aligned} Var\left(\hat{t}_{Y_{AR_o}}\right) \;=\; & \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} \\ & + \boldsymbol{\beta}^t \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}^t \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\gamma}^t \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} + 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L}\boldsymbol{\gamma} - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - \boldsymbol{\gamma}^t \mathbf{L}^t \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} \\ & - 2\ \boldsymbol{\gamma}^t \mathbf{Z}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z}\boldsymbol{\gamma}. \end{aligned}$$

Ricordando che

- $\mathbf{X}^t\ \boldsymbol{\Delta}_{12}\ \mathbf{Z} = \mathbf{0}$.

- $\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{y}$

- $\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}$

- $\mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} = \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{y}^t\ \boldsymbol{\Delta}_2\mathbf{P}\ \mathbf{Z}$

- $\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\mathbf{L} = \mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\left(\mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\right)^{-1} \mathbf{X}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z} = \mathbf{Z}^t\ \boldsymbol{\Delta}_2\mathbf{P}\ \mathbf{Z}$

la (A.28) si riduce

$$\begin{aligned}
Var\left(\hat{t}_{Y_{AR}}\right) &= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} - \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\mathbf{P}\ \mathbf{Z}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&\quad - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{Z}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\mathbf{P}\ \mathbf{Z}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I}-\mathbf{P}\right)\mathbf{Z}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&\quad - 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_2\left(\mathbf{I}-\mathbf{P}\right)\mathbf{Z}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\gamma}^t\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \boldsymbol{\gamma}^t\mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&\quad - 2\ \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} \\
&= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\gamma}^t\left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right)\boldsymbol{\gamma} \\
&\quad - 2\ \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \ 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}
\end{aligned}$$

Dalla (3.12) si ha che

$$\boldsymbol{\gamma}_o^t = \left(\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right)\left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right)^{-1}$$

quindi

$$\begin{aligned}
\boldsymbol{\gamma}_o^t\left(\mathbf{Z}^t\ \mathbf{R}\ \mathbf{Z} + \mathbf{Z}^t\ \boldsymbol{\Delta}_1\ \mathbf{Z}\boldsymbol{\gamma} - \ 2\ \mathbf{Z}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right)\boldsymbol{\gamma} &= \left(\mathbf{y}^t\ \mathbf{R}\ \mathbf{Z} - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\right)\boldsymbol{\gamma} \\
&= \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\ \boldsymbol{\gamma} - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}
\end{aligned}$$

e la varianza da disegno dello stimatore AR_o può essere scritta in maniera semplice come

$$\begin{aligned}
Var\left(\hat{t}_{Y_{AR}}\right) &= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} \\
&\quad - 2\ \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\ \boldsymbol{\gamma} \\
&\quad + 2\ \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma} - \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z} \\
&= \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{y} - \ \mathbf{y}^t\ \boldsymbol{\Delta}_2\ \mathbf{X}\boldsymbol{\beta} - \ \mathbf{y}^t\ \mathbf{R}\ \mathbf{Z}\boldsymbol{\gamma} + \mathbf{y}^t\ \boldsymbol{\Delta}_{21}\ \mathbf{Z}\boldsymbol{\gamma}.
\end{aligned}$$

A9. Risultati dello studio di simulazione

Tutte le possibili combinazioni tra i tre livelli dei coefficienti di regressione parziale

- Alto (H), $0,75 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 1,00$
- Medio (M), $0,40 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 0,60$
- Basso (L), $0,00 \leq r_{(\cdot,\cdot|\cdot)}{}^2 \leq 0,25$

sono stati implementati.

Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$	Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$	Scenario	$r^2_{yx\|z}$	$r^2_{yz\|x}$	$r^2_{xz\|y}$
1	H	H	H	10	M	H	H	19	L	H	H
2	H	H	M	11	M	H	M	20	L	H	M
3	H	H	L	12	M	H	L	21	L	H	L
4	H	M	H	13	M	M	H	22	L	M	H
5	H	M	M	14	M	M	M	23	L	M	M
6	H	M	L	15	M	M	L	24	L	M	L
7	H	L	H	16	M	L	H	25	L	L	H
8	H	L	M	17	M	L	M	26	L	L	M
9	H	L	L	18	M	L	L	27	L	L	L

I risultati per ciascuno scenario sono illustrati di seguito.

SCENARIO 1

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.527	-0.775	-0.113	0.960	0.978	0.945	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$
Case 1							
HT	200,581.1	9.0	35.3	200,050.0	2,255.1	199,303.0	2,255.6
GREGx	200,576.2	6.5	29.8	200,055.8	0.0	199,305.9	2,253.5
GREGxz	200,566.2	1.6	4.6	200,055.8	0.0	199,318.1	0.0
CAL	200,622.9	29.8	61.1	200,055.8	0.0	199,334.7	3,308.3
RW	200,580.4	8.6	58.2	200,055.8	0.0	199,318.1	0.0
AC	200,580.4	8.6	58.2	200,055.8	0.0	199,318.1	0.0
AR	200,566.2	1.6	4.6	200,055.8	0.0	199,318.1	0.0
$\widetilde{AR}$	200,566.2	1.6	4.6	200,055.8	0.0	199,318.1	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199381.113			$sd(\hat{t}_{z_1})$=1591.681		
HT	200,575.8	6.3	34.0	200,110.7	2,165.2	199,268.9	2,186.6
GREGx	200,609.1	23.0	29.0	200,055.8	0.0	199,271.1	2,183.1
GREGxz	200,517.9	-22.5	21.5	200,055.8	0.0	199,381.1	1,591.7
CAL	200,556.0	-3.5	58.6	200,055.8	0.0	199,239.8	3,206.5
RW	200,582.1	9.5	56.3	200,055.8	0.0	199,341.2	1,921.3
AC	200,585.3	11.1	56.0	200,055.8	0.0	199,381.1	1,591.7
AR	200,548.5	-7.3	18.0	200,055.8	0.0	199,344.3	1,324.0
$\widetilde{AR}$	200,548.3	-7.4	17.9	200,055.8	0.0	199,344.5	1,322.7
Case 3		$\mu(\hat{t}_{z_1})$=199328.003			$sd(\hat{t}_{z_1})$=1492.495		
HT	200,581.5	9.2	34.4	199,968.9	2,224.9	199,366.0	2,138.1
GREGx	200,535.3	-13.8	28.4	200,055.8	0.0	199,355.6	2,131.3
GREGxz	200,558.6	-2.2	20.1	200,055.8	0.0	199,328.0	1,492.5
CAL	200,701.3	68.9	60.4	200,055.8	0.0	199,475.2	3,173.3
RW	200,665.6	51.1	59.4	200,055.8	0.0	199,410.7	2,280.4
AC	200,618.7	27.7	54.3	200,055.8	0.0	199,328.0	1,492.5
AR	200,558.6	-2.2	20.1	200,055.8	0.0	199,328.0	1,492.5
$\widetilde{AR}$	200,558.6	-2.2	20.1	200,055.8	0.0	199,328.0	1,492.5
Case 4		$\mu(\hat{t}_{z_1})$=199277.035			$sd(\hat{t}_{z_1})$=1613.246		
HT	200,574.0	5.5	33.3	199,967.1	2,145.0	199,386.7	2,181.2
GREGx	200,521.3	-20.9	28.8	200,055.8	0.0	199,384.1	2,170.4
GREGxz	200,611.9	24.3	21.7	200,055.8	0.0	199,277.0	1,613.2
CAL	200,691.9	64.2	57.6	200,055.8	0.0	199,497.3	3,213.2
RW	200,627.3	32.0	55.2	200,055.8	0.0	199,347.0	1,877.8
AC	200,599.0	17.9	55.5	200,055.8	0.0	199,277.0	1,613.2
AR	200,580.2	8.5	17.7	200,055.8	0.0	199,314.5	1,301.7
$\widetilde{AR}$	200,581.5	9.2	17.6	200,055.8	0.0	199,313.0	1,297.1

SCENARIO 2

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.849	-0.85	-0.528	0.797	0.798	0.478	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$
Case 1							
HT	200,100.7	4.3	35.0	200,123.9	2,237.0	200,174.0	2,226.0
GREGx	200,085.9	-3.1	18.8	200,113.4	0.0	200,189.2	1,932.4
GREGxz	200,106.4	7.1	8.3	200,113.4	0.0	200,155.5	0.0
CAL	200,093.9	0.8	19.6	200,113.4	0.0	200,198.9	3,796.0
RW	200,088.1	-2.0	19.3	200,113.4	0.0	200,155.5	0.0
AC	200,088.1	-2.0	19.3	200,113.4	0.0	200,155.5	0.0
AR	200,106.4	7.1	8.3	200,113.4	0.0	200,155.5	0.0
$\widetilde{AR}$	200,106.4	7.1	8.3	200,113.4	0.0	200,155.5	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200147.474			$sd(\hat{t}_{z_1})$=1583.784		
HT	199,978.7	-56.7	33.4	200,023.2	2,104.8	200,252.0	2,111.7
GREGx	200,051.8	-20.2	17.8	200,113.4	0.0	200,209.1	1,817.4
GREGxz	200,085.1	-3.5	16.1	200,113.4	0.0	200,147.5	1,583.8
CAL	200,071.4	-10.4	18.4	200,113.4	0.0	200,370.1	3,600.5
RW	200,071.5	-10.3	18.2	200,113.4	0.0	200,239.9	2,153.9
AC	200,074.1	-9.0	18.1	200,113.4	0.0	200,147.5	1,583.8
AR	200,070.7	-10.7	13.3	200,113.4	0.0	200,174.0	1,200.3
$\widetilde{AR}$	200,071.7	-10.2	13.3	200,113.4	0.0	200,172.3	1,199.0
Case 3		$\mu(\hat{t}_{z_1})$=200126.382			$sd(\hat{t}_{z_1})$=1563.538		
HT	200,224.0	65.9	35.4	200,288.8	2,259.2	200,058.2	2,204.5
GREGx	200,077.1	-7.5	18.1	200,113.4	0.0	200,150.4	1,856.9
GREGxz	200,088.8	-1.7	16.3	200,113.4	0.0	200,126.4	1,563.5
CAL	200,052.9	-19.6	18.8	200,113.4	0.0	199,925.5	3,840.8
RW	200,044.1	-24.0	18.7	200,113.4	0.0	200,107.6	2,644.9
AC	200,040.2	-26.0	19.1	200,113.4	0.0	200,126.4	1,563.5
AR	200,088.8	-1.7	16.3	200,113.4	0.0	200,126.4	1,563.5
$\widetilde{AR}$	200,088.8	-1.7	16.3	200,113.4	0.0	200,126.4	1,563.5
Case 4		$\mu(\hat{t}_{z_1})$=200076.073			$sd(\hat{t}_{z_1})$=1558.989		
HT	200,142.8	25.3	34.0	200,188.3	2,152.4	200,160.2	2,149.1
GREGx	200,076.1	-8.0	18.0	200,113.4	0.0	200,203.7	1,822.4
GREGxz	200,146.0	26.9	16.5	200,113.4	0.0	200,076.1	1,559.0
CAL	200,071.4	-10.4	18.8	200,113.4	0.0	200,121.0	3,687.3
RW	200,073.1	-9.6	18.6	200,113.4	0.0	200,012.2	2,071.7
AC	200,068.2	-12.0	18.5	200,113.4	0.0	200,076.1	1,559.0
AR	200,117.8	12.8	13.3	200,113.4	0.0	200,127.6	1,137.9
$\widetilde{AR}$	200,117.0	12.4	13.3	200,113.4	0.0	200,128.9	1,135.1

SCENARIO 4

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.896	0.586	0.848	0.864	0.547	0.807	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$
Case 1							
HT	199,918.5	-33.2	33.9	199,689.6	2,183.2	199,369.8	2,244.8
GREGx	199,970.7	-7.1	16.1	199,746.6	0.0	199,414.8	1,218.2
GREGxz	199,968.4	-8.2	10.5	199,746.6	0.0	199,413.4	0.0
CAL	199,978.5	-3.2	16.7	199,746.6	0.0	199,429.4	1,244.4
RW	199,977.5	-3.7	12.8	199,746.6	0.0	199,413.4	0.0
AC	199,977.5	-3.7	12.8	199,746.6	0.0	199,413.4	0.0
AR	199,968.4	-8.2	10.5	199,746.6	0.0	199,413.4	0.0
$\widetilde{AR}$	199,968.4	-8.2	10.5	199,746.6	0.0	199,413.4	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199473.001			$sd(\hat{t}_{z_1})$=1558.952		
HT	200,033.6	24.4	34.9	199,766.3	2,236.4	199,416.8	2,193.4
GREGx	200,017.6	16.4	15.4	199,746.6	0.0	199,396.0	1,191.8
GREGxz	199,966.6	-9.1	18.4	199,746.6	0.0	199,473.0	1,559.0
CAL	200,016.6	15.9	15.9	199,746.6	0.0	199,401.2	1,248.2
RW	199,987.4	1.3	13.8	199,746.6	0.0	199,447.1	716.8
AC	199,971.8	-6.5	17.8	199,746.6	0.0	199,473.0	1,559.0
AR	199,998.1	6.7	13.8	199,746.6	0.0	199,425.5	952.6
$\widetilde{AR}$	199,998.0	6.6	13.8	199,746.6	0.0	199,425.5	954.0
Case 3		$\mu(\hat{t}_{z_1})$=199422.035			$sd(\hat{t}_{z_1})$=1552.436		
HT	199,936.2	-24.3	36.1	199,712.9	2,297.0	199,415.3	2,247.5
GREGx	199,966.9	-8.9	15.6	199,746.6	0.0	199,444.7	1,188.0
GREGxz	199,984.5	-0.2	18.7	199,746.6	0.0	199,422.0	1,552.4
CAL	199,972.4	-6.2	16.0	199,746.6	0.0	199,453.0	1,242.3
RW	199,980.0	-2.4	14.5	199,746.6	0.0	199,433.0	860.0
AC	199,985.1	0.1	21.3	199,746.6	0.0	199,422.0	1,552.4
AR	199,984.5	-0.2	18.7	199,746.6	0.0	199,422.0	1,552.4
$\widetilde{AR}$	199,984.5	-0.2	18.7	199,746.6	0.0	199,422.0	1,552.4
Case 4		$\mu(\hat{t}_{z_1})$=199408.226			$sd(\hat{t}_{z_1})$=1581.991		
HT	199,955.1	-14.8	35.3	199,702.6	2,211.8	199,394.4	2,225.6
GREGx	199,999.4	7.3	16.0	199,746.6	0.0	199,425.7	1,214.1
GREGxz	200,009.7	12.5	18.5	199,746.6	0.0	199,408.2	1,582.0
CAL	200,001.5	8.3	16.4	199,746.6	0.0	199,441.5	1,256.0
RW	200,013.6	14.4	13.6	199,746.6	0.0	199,409.2	694.2
AC	200,015.4	15.3	17.5	199,746.6	0.0	199,408.2	1,582.0
AR	200,004.3	9.8	14.1	199,746.6	0.0	199,417.7	953.6
$\widetilde{AR}$	200,003.2	9.2	14.1	199,746.6	0.0	199,419.5	954.3

SCENARIO 5

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.979	-0.491	0.616	0.971	0.474	0.570	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$
Case 1							
HT	200,159.5	-43.9	35.8	199,826.6	2,271.0	199,911.4	2,248.2
GREGx	200,229.7	-8.9	7.4	199,754.8	0.0	199,866.3	1,787.4
GREGxz	200,228.6	-9.4	5.4	199,754.8	0.0	199,863.6	0.0
CAL	200,138.4	-54.5	69.5	199,754.8	0.0	199,849.9	1,996.2
RW	200,105.9	-70.7	59.9	199,754.8	0.0	199,863.6	0.0
AC	200,105.9	-70.7	59.9	199,754.8	0.0	199,863.6	0.0
AR	200,228.6	-9.4	5.4	199,754.8	0.0	199,863.6	0.0
$\widetilde{AR}$	200,228.6	-9.4	5.4	199,754.8	0.0	199,863.6	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199943.881			$sd(\hat{t}_{z_1})$=1583.555		
HT	200,188.2	-29.6	34.9	199,829.8	2,214.6	199,938.4	2,231.5
GREGx	200,262.7	7.6	7.4	199,754.8	0.0	199,895.1	1,781.1
GREGxz	200,270.5	11.5	6.8	199,754.8	0.0	199,943.9	1,583.6
CAL	200,162.1	-42.6	67.9	199,754.8	0.0	199,873.2	1,973.2
RW	200,187.7	-29.9	61.5	199,754.8	0.0	199,920.6	1,104.4
AC	200,188.7	-29.4	64.8	199,754.8	0.0	199,943.9	1,583.6
AR	200,266.9	9.7	6.2	199,754.8	0.0	199,921.4	1,124.6
$\widetilde{AR}$	200,267.1	9.8	6.2	199,754.8	0.0	199,922.3	1,121.6
Case 3		$\mu(\hat{t}_{z_1})$=199965.976			$sd(\hat{t}_{z_1})$=1499.846		
HT	200,132.5	-57.4	34.4	199,884.4	2,167.5	199,939.8	2,191.4
GREGx	200,259.5	6.0	7.1	199,754.8	0.0	199,858.5	1,798.8
GREGxz	200,277.7	15.1	6.7	199,754.8	0.0	199,966.0	1,499.8
CAL	200,052.3	-97.5	66.7	199,754.8	0.0	199,820.6	2,005.6
RW	200,003.0	-122.1	62.3	199,754.8	0.0	199,804.7	1,396.1
AC	200,165.7	-40.9	47.8	199,754.8	0.0	199,966.0	1,499.8
AR	200,277.7	15.1	6.7	199,754.8	0.0	199,966.0	1,499.8
$\widetilde{AR}$	200,277.7	15.1	6.7	199,754.8	0.0	199,966.0	1,499.8
Case 4		$\mu(\hat{t}_{z_1})$=199877.969			$sd(\hat{t}_{z_1})$=1631.304		
HT	200,218.9	-14.3	33.6	199,786.7	2,119.7	199,920.4	2,172.6
GREGx	200,249.5	1.0	7.2	199,754.8	0.0	199,896.5	1,795.9
GREGxz	200,246.1	-0.7	7.1	199,754.8	0.0	199,878.0	1,631.3
CAL	200,229.9	-8.8	65.2	199,754.8	0.0	199,897.4	1,985.1
RW	200,191.7	-27.9	60.3	199,754.8	0.0	199,885.8	1,117.5
AC	200,158.4	-44.5	65.1	199,754.8	0.0	199,878.0	1,631.3
AR	200,248.0	0.2	6.2	199,754.8	0.0	199,888.3	1,198.4
$\widetilde{AR}$	200,247.7	0.1	6.2	199,754.8	0.0	199,886.7	1,195.4

SCENARIO 6

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.988	0.954	-0.927	0.856	0.445	0.126	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$
Case 1							
HT	200,079.8	-44.6	36.3	199,958.9	2,290.7	200,101.4	2,321.8
GREGx	200,170.2	0.6	5.4	199,867.0	0.0	200,184.0	835.1
GREGxz	200,170.3	0.6	4.1	199,867.0	0.0	200,182.8	0.0
CAL	200,040.6	-64.1	70.6	199,867.0	0.0	200,061.0	4,431.7
RW	200,164.1	-2.4	10.5	199,867.0	0.0	200,182.8	0.0
AC	200,164.1	-2.4	10.5	199,867.0	0.0	200,182.8	0.0
AR	200,170.3	0.6	4.1	199,867.0	0.0	200,182.8	0.0
$\widetilde{AR}$	200,170.3	0.6	4.1	199,867.0	0.0	200,182.8	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200177.929			$sd(\hat{t}_{z_1})$=1536.403		
HT	200,246.7	38.8	35.2	199,789.3	2,262.3	200,268.8	2,203.0
GREGx	200,170.1	0.6	5.4	199,867.0	0.0	200,198.0	839.7
GREGxz	200,165.7	-1.6	7.4	199,867.0	0.0	200,177.9	1,536.4
CAL	200,368.1	99.5	69.2	199,867.0	0.0	200,388.7	4,283.2
RW	200,208.7	19.9	40.2	199,867.0	0.0	200,229.8	2,444.3
AC	200,155.6	-6.7	26.2	199,867.0	0.0	200,177.9	1,536.4
AR	200,169.0	0.0	5.0	199,867.0	0.0	200,193.1	743.1
$\widetilde{AR}$	200,168.8	-0.1	5.0	199,867.0	0.0	200,192.4	742.2
Case 3		$\mu(\hat{t}_{z_1})$=200168.615			$sd(\hat{t}_{z_1})$=1519.396		
HT	200,184.3	7.6	34.1	199,856.3	2,166.2	200,181.0	2,175.9
GREGx	200,174.3	2.6	5.3	199,867.0	0.0	200,170.4	820.2
GREGxz	200,173.9	2.4	7.5	199,867.0	0.0	200,168.6	1,519.4
CAL	200,237.7	34.4	66.6	199,867.0	0.0	200,233.3	4,169.7
RW	200,170.1	0.5	46.8	199,867.0	0.0	200,166.3	2,910.6
AC	200,172.1	1.6	24.5	199,867.0	0.0	200,168.6	1,519.4
AR	200,173.9	2.4	7.5	199,867.0	0.0	200,168.6	1,519.4
$\widetilde{AR}$	200,173.9	2.4	7.5	199,867.0	0.0	200,168.6	1,519.4
Case 4		$\mu(\hat{t}_{z_1})$=200254.996			$sd(\hat{t}_{z_1})$=1495.979		
HT	200,171.6	1.3	35.3	199,881.5	2,227.6	200,195.4	2,216.1
GREGx	200,186.3	8.6	5.2	199,867.0	0.0	200,209.9	827.2
GREGxz	200,196.9	14.0	7.5	199,867.0	0.0	200,255.0	1,496.0
CAL	200,203.4	17.2	68.6	199,867.0	0.0	200,225.6	4,266.3
RW	200,229.6	30.3	38.5	199,867.0	0.0	200,253.3	2,326.4
AC	200,230.4	30.7	26.0	199,867.0	0.0	200,255.0	1,496.0
AR	200,188.8	9.9	4.9	199,867.0	0.0	200,220.2	718.3
$\widetilde{AR}$	200,188.9	10.0	4.9	199,867.0	0.0	200,220.8	717.4

SCENARIO 8

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.946	0.787	-0.896	0.774	0.176	0.572	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,143.8	18.3	34.1	199,767.1	2,163.4	200,400.4	2,171.2
GREGx	200,138.9	15.8	11.1	199,772.1	0.0	200,397.7	977.7
GREGxz	200,120.7	6.7	10.0	199,772.1	0.0	200,446.6	0.0
CAL	200,191.4	42.0	65.8	199,772.1	0.0	200,447.0	4,127.0
RW	200,187.2	40.0	22.6	199,772.1	0.0	200,446.6	0.0
AC	200,187.2	40.0	22.6	199,772.1	0.0	200,446.6	0.0
AR	200,120.7	6.7	10.0	199,772.1	0.0	200,446.6	0.0
$\widetilde{AR}$	200,120.7	6.7	10.0	199,772.1	0.0	200,446.6	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200424.122			$sd(\hat{t}_{z_1})$=1524.949		
HT	200,197.9	45.3	35.1	199,660.9	2,234.7	200,564.9	2,230.2
GREGx	200,090.2	-8.5	11.7	199,772.1	0.0	200,467.6	1,018.7
GREGxz	200,103.3	-2.0	12.7	199,772.1	0.0	200,424.1	1,524.9
CAL	200,349.6	121.1	68.1	199,772.1	0.0	200,716.0	4,258.9
RW	200,135.7	14.2	44.8	199,772.1	0.0	200,497.1	2,448.4
AC	200,063.8	-21.7	34.2	199,772.1	0.0	200,424.1	1,524.9
AR	200,094.1	-6.6	11.2	199,772.1	0.0	200,454.8	825.7
$\widetilde{AR}$	200,094.1	-6.5	11.2	199,772.1	0.0	200,454.6	825.5
Case 3		$\mu(\hat{t}_{z_1})$=200432.493			$sd(\hat{t}_{z_1})$=1590.691		
HT	200,012.0	-47.6	35.1	199,863.9	2,226.7	200,315.1	2,238.3
GREGx	200,098.7	-4.2	11.3	199,772.1	0.0	200,394.3	1,009.3
GREGxz	200,087.7	-9.8	12.7	199,772.1	0.0	200,432.5	1,590.7
CAL	199,970.0	-68.6	67.7	199,772.1	0.0	200,271.8	4,249.8
RW	200,171.9	32.3	50.3	199,772.1	0.0	200,478.4	3,001.6
AC	200,127.1	9.9	30.5	199,772.1	0.0	200,432.5	1,590.7
AR	200,087.7	-9.8	12.7	199,772.1	0.0	200,432.5	1,590.7
$\widetilde{AR}$	200,087.7	-9.8	12.7	199,772.1	0.0	200,432.5	1,590.7
Case 4		$\mu(\hat{t}_{z_1})$=200435.974			$sd(\hat{t}_{z_1})$=1564.106		
HT	200,132.9	12.8	35.8	199,749.4	2,244.0	200,486.3	2,253.1
GREGx	200,112.2	2.5	11.1	199,772.1	0.0	200,466.3	983.0
GREGxz	200,120.8	6.8	12.5	199,772.1	0.0	200,436.0	1,564.1
CAL	200,200.7	46.7	68.8	199,772.1	0.0	200,552.9	4,290.9
RW	200,117.4	5.1	43.9	199,772.1	0.0	200,467.3	2,439.8
AC	200,086.2	-10.5	33.2	199,772.1	0.0	200,436.0	1,564.1
AR	200,114.2	3.5	10.8	199,772.1	0.0	200,459.1	820.4
$\widetilde{AR}$	200,114.4	3.6	10.8	199,772.1	0.0	200,458.6	820.3

SCENARIO 9

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.979	-0.624	-0.647	0.933	0.003	0.05	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,925.3	45.1	35.3	199,951.1	2,231.7	199,747.2	2,270.2
GREGx	199,859.5	12.2	7.3	199,884.3	0.0	199,793.0	1,700.9
GREGxz	199,863.2	14.1	7.3	199,884.3	0.0	199,898.3	0.0
CAL	199,859.0	11.9	7.3	199,884.3	0.0	199,721.3	4,018.6
RW	199,861.7	13.3	7.3	199,884.3	0.0	199,898.3	0.0
AC	199,861.7	13.3	7.3	199,884.3	0.0	199,898.3	0.0
AR	199,863.2	14.1	7.3	199,884.3	0.0	199,898.3	0.0
$\widetilde{AR}$	199,863.2	14.1	7.3	199,884.3	0.0	199,898.3	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199844.264			$sd(\hat{t}_{z_1})$=1547.95		
HT	199,838.6	1.7	35.5	199,894.7	2,237.9	199,898.1	2,257.5
GREGx	199,826.6	-4.3	7.1	199,884.3	0.0	199,903.3	1,687.5
GREGxz	199,825.5	-4.8	7.2	199,884.3	0.0	199,844.3	1,548.0
CAL	199,828.5	-3.3	7.1	199,884.3	0.0	199,926.7	4,029.4
RW	199,827.1	-4.0	7.1	199,884.3	0.0	199,882.7	2,313.3
AC	199,826.1	-4.5	7.1	199,884.3	0.0	199,844.3	1,548.0
AR	199,826.0	-4.6	7.1	199,884.3	0.0	199,872.1	1,136.5
$\widetilde{AR}$	199,826.0	-4.6	7.1	199,884.3	0.0	199,871.9	1,135.1
Case 3		$\mu(\hat{t}_{z_1})$=199945.994			$sd(\hat{t}_{z_1})$=1539.921		
HT	199,859.9	12.4	34.3	199,909.5	2,175.9	199,940.4	2,179.3
GREGx	199,835.8	0.4	7.0	199,884.3	0.0	199,965.2	1,661.1
GREGxz	199,837.6	1.2	7.1	199,884.3	0.0	199,946.0	1,539.9
CAL	199,835.2	0.1	7.1	199,884.3	0.0	199,952.6	3,873.2
RW	199,837.1	1.0	7.0	199,884.3	0.0	200,002.5	2,758.2
AC	199,835.8	0.4	7.0	199,884.3	0.0	199,946.0	1,539.9
AR	199,837.6	1.2	7.1	199,884.3	0.0	199,946.0	1,539.9
$\widetilde{AR}$	199,837.6	1.2	7.1	199,884.3	0.0	199,946.0	1,539.9
Case 4		$\mu(\hat{t}_{z_1})$=199831.898			$sd(\hat{t}_{z_1})$=1514.952		
HT	199,816.1	-9.5	36.3	199,880.0	2,299.5	199,959.4	2,235.2
GREGx	199,820.3	-7.4	7.2	199,884.3	0.0	199,958.1	1,707.7
GREGxz	199,817.4	-8.9	7.3	199,884.3	0.0	199,831.9	1,515.0
CAL	199,820.9	-7.1	7.2	199,884.3	0.0	200,003.6	4,032.3
RW	199,820.1	-7.5	7.2	199,884.3	0.0	199,887.5	2,223.5
AC	199,819.3	-7.9	7.2	199,884.3	0.0	199,831.9	1,515.0
AR	199,818.8	-8.2	7.2	199,884.3	0.0	199,890.1	1,104.5
$\widetilde{AR}$	199,818.7	-8.2	7.2	199,884.3	0.0	199,890.5	1,105.8

SCENARIO 10

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.136	0.779	-0.641	0.571	0.828	0.743	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,031.6	1.2	34.3	200,215.6	2,149.7	199,858.4	2,228.6
GREGx	200,044.1	7.4	33.8	200,160.9	0.0	199,894.8	1,688.3
GREGxz	199,992.4	-18.4	14.2	200,160.9	0.0	199,848.1	0.0
CAL	200,004.0	-12.6	51.8	200,160.9	0.0	199,841.7	3,900.5
RW	200,013.0	-8.1	17.6	200,160.9	0.0	199,848.1	0.0
AC	200,013.0	-8.1	17.6	200,160.9	0.0	199,848.1	0.0
AR	199,992.4	-18.4	14.2	200,160.9	0.0	199,848.1	0.0
$\widetilde{AR}$	199,992.4	-18.4	14.2	200,160.9	0.0	199,848.1	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199873.937			$sd(\hat{t}_{z_1})$=1529.069		
HT	200,102.7	36.8	35.9	200,047.6	2,208.9	199,990.0	2,157.9
GREGx	200,089.2	30.0	36.2	200,160.9	0.0	199,922.3	1,752.6
GREGxz	200,037.1	3.9	30.9	200,160.9	0.0	199,873.9	1,529.1
CAL	200,237.4	104.1	51.4	200,160.9	0.0	200,134.6	3,823.0
RW	200,123.2	47.0	32.2	200,160.9	0.0	199,986.9	2,198.4
AC	200,036.5	3.6	25.7	200,160.9	0.0	199,873.9	1,529.1
AR	200,056.8	13.8	25.1	200,160.9	0.0	199,892.6	1,129.1
$\widetilde{AR}$	200,058.4	14.6	25.1	200,160.9	0.0	199,893.9	1,127.6
Case 3		$\mu(\hat{t}_{z_1})$=199771.752			$sd(\hat{t}_{z_1})$=1545.521		
HT	200,011.1	-9.1	34.9	200,288.7	2,264.5	199,731.9	2,174.2
GREGx	200,030.3	0.5	35.0	200,160.9	0.0	199,813.5	1,706.6
GREGxz	199,981.0	-24.1	32.4	200,160.9	0.0	199,771.8	1,545.5
CAL	199,912.4	-58.4	51.6	200,160.9	0.0	199,646.7	3,914.8
RW	199,932.7	-48.2	38.6	200,160.9	0.0	199,669.0	2,758.8
AC	200,015.9	-6.7	29.0	200,160.9	0.0	199,771.8	1,545.5
AR	199,981.0	-24.1	32.4	200,160.9	0.0	199,771.8	1,545.5
$\widetilde{AR}$	199,981.0	-24.1	32.4	200,160.9	0.0	199,771.8	1,545.5
Case 4		$\mu(\hat{t}_{z_1})$=199836.515			$sd(\hat{t}_{z_1})$=1584.302		
HT	200,015.4	-6.9	33.7	200,096.2	2,267.1	199,899.3	2,206.5
GREGx	200,006.3	-11.5	33.3	200,160.9	0.0	199,856.2	1,620.9
GREGxz	199,985.6	-21.8	32.6	200,160.9	0.0	199,836.5	1,584.3
CAL	200,105.5	38.1	52.3	200,160.9	0.0	200,000.5	4,012.8
RW	199,991.1	-19.1	32.8	200,160.9	0.0	199,850.7	2,268.2
AC	199,981.4	-23.9	26.5	200,160.9	0.0	199,836.5	1,584.3
AR	199,989.7	-19.7	25.2	200,160.9	0.0	199,840.9	1,131.0
$\widetilde{AR}$	199,992.4	-18.4	25.2	200,160.9	0.0	199,843.1	1,130.8

SCENARIO 11

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.245	-0.929	0.002	0.444	0.919	0.408	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,819.4	-34.8	34.2	199,524.9	2,192.4	200,046.4	2,184.4
GREGx	199,851.1	-19.0	33.5	199,639.1	0.0	200,040.2	2,193.9
GREGxz	199,888.4	-0.3	10.0	199,639.1	0.0	200,002.0	0.0
CAL	199,949.7	30.4	42.9	199,639.1	0.0	200,179.6	2,997.8
RW	199,938.3	24.7	42.8	199,639.1	0.0	200,002.0	0.0
AC	199,938.3	24.7	42.8	199,639.1	0.0	200,002.0	0.0
AR	199,888.4	-0.3	10.0	199,639.1	0.0	200,002.0	0.0
$\widetilde{AR}$	199,888.4	-0.3	10.0	199,639.1	0.0	200,002.0	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200031.686			$sd(\hat{t}_{z_1})$=1480.706		
HT	199,807.4	-40.8	34.3	199,626.8	2,272.9	200,100.5	2,216.9
GREGx	199,807.8	-40.6	33.8	199,639.1	0.0	200,103.6	2,227.8
GREGxz	199,876.9	-6.1	24.3	199,639.1	0.0	200,031.7	1,480.7
CAL	199,838.7	-25.2	43.7	199,639.1	0.0	200,135.6	3,071.8
RW	199,831.9	-28.6	43.7	199,639.1	0.0	200,070.3	1,730.2
AC	199,830.5	-29.3	43.7	199,639.1	0.0	200,031.7	1,480.7
AR	199,853.5	-17.8	20.8	199,639.1	0.0	200,056.1	1,247.4
$\widetilde{AR}$	199,854.0	-17.5	20.8	199,639.1	0.0	200,055.6	1,245.6
Case 3		$\mu(\hat{t}_{z_1})$=200099.959			$sd(\hat{t}_{z_1})$=2267.852		
HT	199,816.3	-36.4	35.4	199,722.7	2,173.0	200,100.0	2,267.9
GREGx	199,803.7	-42.7	34.6	199,639.1	0.0	200,091.5	2,272.1
GREGxz	199,795.9	-46.6	34.6	199,639.1	0.0	200,100.0	2,267.9
CAL	199,751.1	-69.0	43.1	199,639.1	0.0	200,039.9	3,118.9
RW	199,751.1	-69.0	43.1	199,639.1	0.0	200,039.9	3,118.9
AC	199,727.2	-81.0	45.3	199,639.1	0.0	200,100.0	2,267.9
AR	199,795.9	-46.6	34.6	199,639.1	0.0	200,100.0	2,267.9
$\widetilde{AR}$	199,795.9	-46.6	34.6	199,639.1	0.0	200,100.0	2,267.9
Case 4		$\mu(\hat{t}_{z_1})$=200114.108			$sd(\hat{t}_{z_1})$=2228.7		
HT	199,957.2	34.1	34.3	199,681.4	2,284.7	199,918.7	2,155.2
GREGx	199,943.3	27.1	33.3	199,639.1	0.0	199,918.6	2,163.0
GREGxz	199,762.0	-63.5	35.2	199,639.1	0.0	200,114.1	2,228.7
CAL	199,933.8	22.4	42.9	199,639.1	0.0	199,900.5	3,077.1
RW	199,908.7	9.8	42.8	199,639.1	0.0	199,966.2	2,186.4
AC	199,881.9	-3.6	42.9	199,639.1	0.0	200,114.1	2,228.7
AR	199,857.1	-16.0	24.8	199,639.1	0.0	200,011.4	1,506.0
$\widetilde{AR}$	199,853.2	-18.0	24.8	199,639.1	0.0	200,015.8	1,501.8

SCENARIO 12

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.732	-0.89	-0.499	0.534	0.792	0.245	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,103.3	63.6	33.6	200,209.4	2,210.7	200,006.5	2,107.5
GREGx	199,953.6	-11.2	23.4	200,004.7	0.0	200,105.7	1,861.3
GREGxz	199,973.3	-1.4	11.0	200,004.7	0.0	200,079.7	0.0
CAL	199,907.1	-34.5	25.5	200,004.7	0.0	199,842.6	3,639.8
RW	199,885.8	-45.1	25.5	200,004.7	0.0	200,079.7	0.0
AC	199,885.8	-45.1	25.5	200,004.7	0.0	200,079.7	0.0
AR	199,973.3	-1.4	11.0	200,004.7	0.0	200,079.7	0.0
$\widetilde{AR}$	199,973.3	-1.4	11.0	200,004.7	0.0	200,079.7	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200077.004			$sd(\hat{t}_{z_1})$=1604.404		
HT	200,036.0	29.9	34.9	200,099.8	2,133.7	200,083.1	2,279.9
GREGx	199,961.7	-7.2	24.5	200,004.7	0.0	200,137.6	1,990.8
GREGxz	200,006.6	15.3	21.4	200,004.7	0.0	200,077.0	1,604.4
CAL	199,947.0	-14.5	25.8	200,004.7	0.0	200,024.5	3,744.8
RW	199,944.7	-15.7	25.8	200,004.7	0.0	200,038.9	2,183.7
AC	199,942.1	-17.0	25.8	200,004.7	0.0	200,077.0	1,604.4
AR	199,987.5	5.7	18.0	200,004.7	0.0	200,102.9	1,265.8
$\widetilde{AR}$	199,989.2	6.6	17.9	200,004.7	0.0	200,100.5	1,260.4
Case 3		$\mu(\hat{t}_{z_1})$=200105.571			$sd(\hat{t}_{z_1})$=1534.721		
HT	200,000.7	12.3	35.2	200,075.4	2,236.0	200,120.4	2,226.4
GREGx	199,947.3	-14.4	24.3	200,004.7	0.0	200,156.4	1,942.7
GREGxz	199,985.2	4.6	20.4	200,004.7	0.0	200,105.6	1,534.7
CAL	199,937.0	-19.5	26.0	200,004.7	0.0	200,087.1	3,786.8
RW	199,930.9	-22.6	26.0	200,004.7	0.0	200,067.4	2,604.5
AC	199,925.8	-25.1	26.3	200,004.7	0.0	200,105.6	1,534.7
AR	199,985.2	4.6	20.4	200,004.7	0.0	200,105.6	1,534.7
$\widetilde{AR}$	199,985.2	4.6	20.4	200,004.7	0.0	200,105.6	1,534.7
Case 4		$\mu(\hat{t}_{z_1})$=200080.797			$sd(\hat{t}_{z_1})$=1527.871		
HT	200,008.0	16.0	33.9	200,046.1	2,156.9	200,050.3	2,052.2
GREGx	199,970.3	-2.9	23.2	200,004.7	0.0	200,074.8	1,787.4
GREGxz	199,964.7	-5.7	20.9	200,004.7	0.0	200,080.8	1,527.9
CAL	199,972.6	-1.7	24.9	200,004.7	0.0	200,042.4	3,577.1
RW	199,965.0	-5.5	24.8	200,004.7	0.0	200,043.7	2,119.0
AC	199,960.9	-7.6	24.9	200,004.7	0.0	200,080.8	1,527.9
AR	199,967.2	-4.4	17.8	200,004.7	0.0	200,078.0	1,166.3
$\widetilde{AR}$	199,965.5	-5.3	17.7	200,004.7	0.0	200,080.4	1,160.5

SCENARIO 13

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.178	-0.364	0.751	0.538	0.586	0.792	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,039.9	24.7	36.8	200,260.1	2,222.4	200,589.5	2,195.7
GREGx	200,014.3	11.9	36.1	200,094.3	0.0	200,462.5	1,491.7
GREGxz	200,052.4	31.0	22.7	200,094.3	0.0	200,430.3	0.0
CAL	199,895.2	-47.6	44.6	200,094.3	0.0	200,430.6	1,598.9
RW	199,868.3	-61.1	41.8	200,094.3	0.0	200,430.3	0.0
AC	199,868.3	-61.1	41.8	200,094.3	0.0	200,430.3	0.0
AR	200,052.4	31.0	22.7	200,094.3	0.0	200,430.3	0.0
$\widetilde{AR}$	200,052.4	31.0	22.7	200,094.3	0.0	200,430.3	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200402.389			$sd(\hat{t}_{z_1})$=1513.774		
HT	199,886.0	-52.2	34.8	200,109.9	2,234.7	200,479.7	2,184.8
GREGx	199,879.5	-55.5	34.2	200,094.3	0.0	200,468.7	1,405.9
GREGxz	199,959.6	-15.4	35.7	200,094.3	0.0	200,402.4	1,513.8
CAL	199,888.9	-50.8	43.6	200,094.3	0.0	200,470.0	1,504.0
RW	199,906.0	-42.3	42.5	200,094.3	0.0	200,413.0	873.6
AC	199,917.9	-36.3	44.0	200,094.3	0.0	200,402.4	1,513.8
AR	199,921.9	-34.3	29.4	200,094.3	0.0	200,433.4	1,020.5
$\widetilde{AR}$	199,921.2	-34.6	29.3	200,094.3	0.0	200,434.1	1,020.4
Case 3		$\mu(\hat{t}_{z_1})$=200444.614			$sd(\hat{t}_{z_1})$=2281.744		
HT	199,930.7	-29.9	36.0	200,092.5	2,205.5	200,444.6	2,281.7
GREGx	199,942.3	-24.1	35.8	200,094.3	0.0	200,438.0	1,493.1
GREGxz	199,933.8	-28.3	47.9	200,094.3	0.0	200,444.6	2,281.7
CAL	199,951.9	-19.3	46.1	200,094.3	0.0	200,451.0	1,561.9
RW	199,951.9	-19.3	46.1	200,094.3	0.0	200,451.0	1,561.9
AC	199,918.9	-35.8	62.4	200,094.3	0.0	200,444.6	2,281.7
AR	199,933.8	-28.3	47.9	200,094.3	0.0	200,444.6	2,281.7
$\widetilde{AR}$	199,933.8	-28.3	47.9	200,094.3	0.0	200,444.6	2,281.7
Case 4		$\mu(\hat{t}_{z_1})$=200360.548			$sd(\hat{t}_{z_1})$=2183.388		
HT	199,955.3	-17.6	34.5	200,082.0	2,187.2	200,423.5	2,240.3
GREGx	199,954.3	-18.1	34.1	200,094.3	0.0	200,436.6	1,454.1
GREGxz	200,041.1	25.3	46.4	200,094.3	0.0	200,360.5	2,183.4
CAL	199,986.2	-2.1	44.6	200,094.3	0.0	200,440.7	1,522.0
RW	200,005.4	7.5	43.8	200,094.3	0.0	200,400.4	1,080.3
AC	200,024.3	16.9	47.3	200,094.3	0.0	200,360.5	2,183.4
AR	199,980.3	-5.1	31.5	200,094.3	0.0	200,414.2	1,226.4
$\widetilde{AR}$	199,979.8	-5.3	31.4	200,094.3	0.0	200,414.3	1,226.3

SCENARIO 14

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.358	0.46	0.371	0.412	0.469	0.419	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,006.2	-26.1	33.2	200,012.2	2,194.3	200,060.4	2,225.4
GREGx	200,034.1	-12.2	31.1	199,911.2	0.0	200,018.9	2,052.8
GREGxz	200,001.2	-28.6	22.5	199,911.2	0.0	199,971.4	0.0
CAL	199,938.3	-60.1	55.2	199,911.2	0.0	199,974.3	2,427.3
RW	199,913.7	-72.4	35.2	199,911.2	0.0	199,971.4	0.0
AC	199,913.7	-72.4	35.2	199,911.2	0.0	199,971.4	0.0
AR	200,001.2	-28.6	22.5	199,911.2	0.0	199,971.4	0.0
$\widetilde{AR}$	200,001.2	-28.6	22.5	199,911.2	0.0	199,971.4	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199909.088			$sd(\hat{t}_{z_1})$=1541.848		
HT	200,028.9	-14.8	35.2	199,821.2	2,238.6	199,857.2	2,300.9
GREGx	199,982.1	-38.2	32.7	199,911.2	0.0	199,880.2	2,126.1
GREGxz	200,010.3	-24.1	29.4	199,911.2	0.0	199,909.1	1,541.8
CAL	200,147.9	44.7	57.6	199,911.2	0.0	199,959.5	2,490.9
RW	200,117.4	29.5	45.4	199,911.2	0.0	199,940.0	1,420.6
AC	200,083.7	12.6	46.9	199,911.2	0.0	199,909.1	1,541.8
AR	200,001.3	-28.6	27.5	199,911.2	0.0	199,900.8	1,253.9
$\widetilde{AR}$	200,000.9	-28.8	27.5	199,911.2	0.0	199,899.9	1,251.5
Case 3		$\mu(\hat{t}_{z_1})$=200025.76			$sd(\hat{t}_{z_1})$=2232.197		
HT	200,049.1	-4.7	35.3	199,810.1	2,311.5	200,025.8	2,232.2
GREGx	200,006.0	-26.2	32.5	199,911.2	0.0	200,060.8	2,056.9
GREGxz	199,982.4	-38.0	33.6	199,911.2	0.0	200,025.8	2,232.2
CAL	200,181.4	61.4	59.2	199,911.2	0.0	200,140.7	2,475.9
RW	200,181.4	61.4	59.2	199,911.2	0.0	200,140.7	2,475.9
AC	200,033.9	-12.3	33.5	199,911.2	0.0	200,025.8	2,232.2
AR	199,982.4	-38.0	33.6	199,911.2	0.0	200,025.8	2,232.2
$\widetilde{AR}$	199,982.4	-38.0	33.6	199,911.2	0.0	200,025.8	2,232.2
Case 4		$\mu(\hat{t}_{z_1})$=199994.695			$sd(\hat{t}_{z_1})$=2247.051		
HT	200,205.6	73.5	35.8	199,820.9	2,144.3	200,037.2	2,364.7
GREGx	200,174.9	58.2	33.7	199,911.2	0.0	200,079.5	2,199.8
GREGxz	200,127.3	34.4	34.3	199,911.2	0.0	199,994.7	2,247.1
CAL	200,321.9	131.7	56.2	199,911.2	0.0	200,139.0	2,518.6
RW	200,212.7	77.1	46.5	199,911.2	0.0	200,048.5	1,751.5
AC	200,155.0	48.2	55.9	199,911.2	0.0	199,994.7	2,247.1
AR	200,154.4	47.9	29.2	199,911.2	0.0	200,042.2	1,540.3
$\widetilde{AR}$	200,151.7	46.6	29.2	199,911.2	0.0	200,038.4	1,530.2

SCENARIO 15

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.671	-0.492	0.029	0.570	0.407	0.218	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,413.7	53.9	35.7	200,241.0	2,239.3	200,546.5	2,233.4
GREGx	199,411.3	52.7	26.4	200,251.9	0.0	200,540.4	2,237.8
GREGxz	199,363.3	28.6	19.8	200,251.9	0.0	200,639.3	0.0
CAL	199,462.7	78.5	63.8	200,251.9	0.0	200,579.3	3,080.8
RW	199,460.4	77.4	57.8	200,251.9	0.0	200,639.3	0.0
AC	199,460.4	77.4	57.8	200,251.9	0.0	200,639.3	0.0
AR	199,363.3	28.6	19.8	200,251.9	0.0	200,639.3	0.0
$\widetilde{AR}$	199,363.3	28.6	19.8	200,251.9	0.0	200,639.3	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200694.417			$sd(\hat{t}_{z_1})$=1574.927		
HT	199,308.6	1.2	34.9	200,267.5	2,178.7	200,567.1	2,259.7
GREGx	199,311.4	2.6	25.9	200,251.9	0.0	200,575.1	2,274.9
GREGxz	199,253.5	-26.5	23.4	200,251.9	0.0	200,694.4	1,574.9
CAL	199,330.2	12.0	62.1	200,251.9	0.0	200,574.0	3,117.2
RW	199,304.8	-0.7	58.0	200,251.9	0.0	200,568.5	1,706.0
AC	199,367.1	30.5	57.7	200,251.9	0.0	200,694.4	1,574.9
AR	199,272.0	-17.2	22.2	200,251.9	0.0	200,656.4	1,302.7
$\widetilde{AR}$	199,272.8	-16.8	22.2	200,251.9	0.0	200,654.7	1,300.1
Case 3		$\mu(\hat{t}_{z_1})$=200548.466			$sd(\hat{t}_{z_1})$=2210.629		
HT	199,309.0	1.4	33.7	200,242.9	2,191.6	200,548.5	2,210.6
GREGx	199,308.9	1.3	25.5	200,251.9	0.0	200,540.1	2,219.8
GREGxz	199,304.9	-0.7	25.4	200,251.9	0.0	200,548.5	2,210.6
CAL	199,353.7	23.8	61.0	200,251.9	0.0	200,579.3	3,092.4
RW	199,353.7	23.8	61.0	200,251.9	0.0	200,579.3	3,092.4
AC	199,297.6	-4.3	44.1	200,251.9	0.0	200,548.5	2,210.6
AR	199,304.9	-0.7	25.4	200,251.9	0.0	200,548.5	2,210.6
$\widetilde{AR}$	199,304.9	-0.7	25.4	200,251.9	0.0	200,548.5	2,210.6
Case 4		$\mu(\hat{t}_{z_1})$=200639.753			$sd(\hat{t}_{z_1})$=2275.152		
HT	199,286.6	-9.8	34.3	200,218.5	2,112.2	200,748.9	2,146.9
GREGx	199,260.0	-23.2	25.8	200,251.9	0.0	200,755.1	2,155.2
GREGxz	199,315.4	4.6	25.8	200,251.9	0.0	200,639.8	2,275.2
CAL	199,352.8	23.3	60.5	200,251.9	0.0	200,801.6	2,939.4
RW	199,278.2	-14.1	58.0	200,251.9	0.0	200,687.0	2,132.5
AC	199,244.6	-30.9	59.2	200,251.9	0.0	200,639.8	2,275.2
AR	199,291.0	-7.6	23.0	200,251.9	0.0	200,690.4	1,582.0
$\widetilde{AR}$	199,287.6	-9.4	23.0	200,251.9	0.0	200,697.6	1,576.9

SCENARIO 16

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.579	0.289	-0.854	0.443	0.233	0.773	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,592.6	35.5	34.2	200,108.3	2,135.9	200,078.4	2,139.2
GREGx	199,527.9	3.1	28.3	200,223.3	0.0	199,979.1	1,094.0
GREGxz	199,543.3	10.8	25.3	200,223.3	0.0	199,963.1	0.0
CAL	199,736.5	107.6	59.1	200,223.3	0.0	200,228.1	4,032.5
RW	199,544.9	11.6	36.7	200,223.3	0.0	199,963.1	0.0
AC	199,544.9	11.6	36.7	200,223.3	0.0	199,963.1	0.0
AR	199,543.3	10.8	25.3	200,223.3	0.0	199,963.1	0.0
$\widetilde{AR}$	199,543.3	10.8	25.3	200,223.3	0.0	199,963.1	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199921.889			$sd(\hat{t}_{z_1})$=1563.536		
HT	199,540.0	9.1	34.6	200,126.3	2,250.4	200,077.1	2,208.2
GREGx	199,483.0	-19.4	28.0	200,223.3	0.0	199,992.5	1,167.3
GREGxz	199,540.9	9.6	30.9	200,223.3	0.0	199,921.9	1,563.5
CAL	199,669.8	74.2	61.6	200,223.3	0.0	200,212.5	4,193.4
RW	199,515.0	-3.4	46.3	200,223.3	0.0	200,000.7	2,387.8
AC	199,456.9	-32.5	41.2	200,223.3	0.0	199,921.9	1,563.5
AR	199,500.7	-10.6	26.9	200,223.3	0.0	199,971.4	949.9
$\widetilde{AR}$	199,501.5	-10.2	26.9	200,223.3	0.0	199,970.4	949.5
Case 3		$\mu(\hat{t}_{z_1})$=200067.081			$sd(\hat{t}_{z_1})$=2277.397		
HT	199,541.9	10.1	35.0	200,118.1	2,282.6	200,067.1	2,277.4
GREGx	199,476.9	-22.5	28.8	200,223.3	0.0	199,974.2	1,163.0
GREGxz	199,408.9	-56.6	37.3	200,223.3	0.0	200,067.1	2,277.4
CAL	199,680.7	79.6	61.9	200,223.3	0.0	200,212.9	4,301.2
RW	199,680.7	79.6	61.9	200,223.3	0.0	200,212.9	4,301.2
AC	199,568.0	23.1	41.3	200,223.3	0.0	200,067.1	2,277.4
AR	199,408.9	-56.6	37.3	200,223.3	0.0	200,067.1	2,277.4
$\widetilde{AR}$	199,408.9	-56.6	37.3	200,223.3	0.0	200,067.1	2,277.4
Case 4		$\mu(\hat{t}_{z_1})$=200041.881			$sd(\hat{t}_{z_1})$=2188.266		
HT	199,483.3	-19.3	34.7	200,306.0	2,185.7	199,891.8	2,206.9
GREGx	199,530.2	4.2	28.8	200,223.3	0.0	199,958.2	1,160.1
GREGxz	199,470.3	-25.8	36.4	200,223.3	0.0	200,041.9	2,188.3
CAL	199,439.0	-41.5	60.0	200,223.3	0.0	199,854.5	4,130.1
RW	199,535.8	7.0	49.7	200,223.3	0.0	199,988.5	2,886.0
AC	199,573.8	26.1	45.3	200,223.3	0.0	200,041.9	2,188.3
AR	199,519.8	-1.0	28.2	200,223.3	0.0	199,973.6	1,021.3
$\widetilde{AR}$	199,287.6	-9.4	23.0	200,251.9	0.0	200,697.6	1,576.9

SCENARIO 17

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.686	-0.242	0.684	0.540	0.183	0.538	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,689.7	-36.0	33.5	200,224.1	2,202.4	200,207.0	2,266.0
GREGx	199,683.0	-39.3	24.2	200,224.8	0.0	200,201.9	1,589.4
GREGxz	199,670.3	-45.7	22.1	200,224.8	0.0	200,175.4	0.0
CAL	199,727.7	-17.0	61.5	200,224.8	0.0	200,213.4	1,682.4
RW	199,656.2	-52.8	54.6	200,224.8	0.0	200,175.4	0.0
AC	199,656.2	-52.8	54.6	200,224.8	0.0	200,175.4	0.0
AR	199,670.3	-45.7	22.1	200,224.8	0.0	200,175.4	0.0
$\widetilde{AR}$	199,670.3	-45.7	22.1	200,224.8	0.0	200,175.4	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200180.719			$sd(\hat{t}_{z_1})$=1525.782		
HT	199,688.2	-36.7	36.0	200,191.9	2,304.1	200,084.3	2,238.6
GREGx	199,674.9	-43.4	25.5	200,224.8	0.0	200,107.6	1,629.1
GREGxz	199,704.1	-28.8	25.0	200,224.8	0.0	200,180.7	1,525.8
CAL	199,761.4	-0.1	65.5	200,224.8	0.0	200,125.1	1,792.2
RW	199,816.3	27.4	58.9	200,224.8	0.0	200,189.1	1,033.6
AC	199,798.9	18.7	62.6	200,224.8	0.0	200,180.7	1,525.8
AR	199,690.0	-35.9	24.1	200,224.8	0.0	200,144.8	1,113.7
$\widetilde{AR}$	199,689.7	-36.0	24.1	200,224.8	0.0	200,144.9	1,115.1
Case 3		$\mu(\hat{t}_{z_1})$=200218.625			$sd(\hat{t}_{z_1})$=2200.842		
HT	199,787.9	13.1	34.0	200,187.3	2,149.3	200,218.6	2,200.8
GREGx	199,767.6	3.0	25.6	200,224.8	0.0	200,248.0	1,679.3
GREGxz	199,751.2	-5.2	27.8	200,224.8	0.0	200,218.6	2,200.8
CAL	199,859.3	48.9	60.7	200,224.8	0.0	200,263.1	1,821.0
RW	199,859.3	48.9	60.7	200,224.8	0.0	200,263.1	1,821.0
AC	199,787.0	12.7	29.8	200,224.8	0.0	200,218.6	2,200.8
AR	199,751.2	-5.2	27.8	200,224.8	0.0	200,218.6	2,200.8
$\widetilde{AR}$	199,751.2	-5.2	27.8	200,224.8	0.0	200,218.6	2,200.8
Case 4		$\mu(\hat{t}_{z_1})$=200212.038			$sd(\hat{t}_{z_1})$=2215.166		
HT	199,786.8	12.6	35.0	200,121.3	2,257.3	200,178.4	2,246.4
GREGx	199,719.1	-21.3	26.7	200,224.8	0.0	200,246.2	1,669.4
GREGxz	199,699.7	-31.0	29.3	200,224.8	0.0	200,212.0	2,215.2
CAL	199,924.8	81.7	63.0	200,224.8	0.0	200,288.9	1,818.7
RW	199,835.0	36.7	58.0	200,224.8	0.0	200,229.2	1,221.0
AC	199,788.1	13.2	69.3	200,224.8	0.0	200,212.0	2,215.2
AR	199,712.5	-24.6	26.0	200,224.8	0.0	200,234.8	1,306.5
$\widetilde{AR}$	199,712.1	-24.8	26.0	200,224.8	0.0	200,233.2	1,304.1

SCENARIO 18

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.624	0.073	-0.14	0.413	0.043	0.057	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,453.4	-34.4	33.5	200,290.1	2,190.1	199,662.4	2,205.9
GREGx	200,517.5	-2.5	26.6	200,395.5	0.0	199,650.5	2,186.5
GREGxz	200,508.3	-7.1	26.1	200,395.5	0.0	199,563.7	0.0
CAL	200,567.0	22.2	29.9	200,395.5	0.0	199,790.2	3,284.0
RW	200,510.1	-6.2	26.5	200,395.5	0.0	199,563.7	0.0
AC	200,510.1	-6.2	26.5	200,395.5	0.0	199,563.7	0.0
AR	200,508.3	-7.1	26.1	200,395.5	0.0	199,563.7	0.0
$\widetilde{AR}$	200,508.3	-7.1	26.1	200,395.5	0.0	199,563.7	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199599.876			$sd(\hat{t}_{z_1})$=1576.317		
HT	200,565.5	21.4	34.4	200,378.7	2,196.6	199,476.6	2,295.4
GREGx	200,575.1	26.2	27.1	200,395.5	0.0	199,479.2	2,282.6
GREGxz	200,595.3	36.3	27.0	200,395.5	0.0	199,599.9	1,576.3
CAL	200,590.9	34.1	30.2	200,395.5	0.0	199,518.6	3,334.2
RW	200,603.7	40.5	28.2	200,395.5	0.0	199,583.3	1,844.1
AC	200,606.8	42.1	27.8	200,395.5	0.0	199,599.9	1,576.3
AR	200,588.4	32.9	26.8	200,395.5	0.0	199,560.6	1,304.3
$\widetilde{AR}$	200,588.4	32.9	26.8	200,395.5	0.0	199,559.3	1,301.7
Case 3		$\mu(\hat{t}_{z_1})$=199693.506			$sd(\hat{t}_{z_1})$=2199.36		
HT	200,437.7	-42.3	34.5	200,305.8	2,098.3	199,693.5	2,199.4
GREGx	200,493.3	-14.5	27.6	200,395.5	0.0	199,677.6	2,184.0
GREGxz	200,495.8	-13.3	27.6	200,395.5	0.0	199,693.5	2,199.4
CAL	200,534.3	5.9	30.1	200,395.5	0.0	199,803.8	3,195.9
RW	200,534.3	5.9	30.1	200,395.5	0.0	199,803.8	3,195.9
AC	200,504.3	-9.1	27.9	200,395.5	0.0	199,693.5	2,199.4
AR	200,495.8	-13.3	27.6	200,395.5	0.0	199,693.5	2,199.4
$\widetilde{AR}$	200,495.8	-13.3	27.6	200,395.5	0.0	199,693.5	2,199.4
Case 4		$\mu(\hat{t}_{z_1})$=199674.157			$sd(\hat{t}_{z_1})$=2324.83		
HT	200,522.7	0.1	34.8	200,491.5	2,176.8	199,513.9	2,212.8
GREGx	200,462.9	-29.7	27.1	200,395.5	0.0	199,525.3	2,200.0
GREGxz	200,482.0	-20.2	26.9	200,395.5	0.0	199,674.2	2,324.8
CAL	200,435.4	-43.4	29.4	200,395.5	0.0	199,445.5	3,236.6
RW	200,479.2	-21.6	28.0	200,395.5	0.0	199,626.9	2,324.6
AC	200,488.1	-17.1	28.0	200,395.5	0.0	199,674.2	2,324.8
AR	200,472.8	-24.8	26.6	200,395.5	0.0	199,601.1	1,592.4
$\widetilde{AR}$	200,472.4	-25.0	26.6	200,395.5	0.0	199,599.0	1,583.1

SCENARIO 20

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.394	0.883	-0.64	0.227	0.799	0.46	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,186.7	34.9	34.7	200,311.4	2,225.1	199,802.2	2,237.4
GREGx	200,207.9	45.5	31.8	200,282.5	0.0	199,826.5	1,711.2
GREGxz	200,148.0	15.6	14.3	200,282.5	0.0	199,770.7	0.0
CAL	200,191.1	37.1	57.5	200,282.5	0.0	199,812.4	3,967.9
RW	200,159.9	21.5	15.1	200,282.5	0.0	199,770.7	0.0
AC	200,159.9	21.5	15.1	200,282.5	0.0	199,770.7	0.0
AR	200,148.0	15.6	14.3	200,282.5	0.0	199,770.7	0.0
$\widetilde{AR}$	200,148.0	15.6	14.3	200,282.5	0.0	199,770.7	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199789.758			$sd(\hat{t}_{z_1})$=1583.701		
HT	200,111.2	-2.8	34.2	200,283.0	2,228.4	199,782.3	2,158.9
GREGx	200,111.4	-2.7	31.9	200,282.5	0.0	199,783.0	1,696.1
GREGxz	200,116.8	0.0	29.4	200,282.5	0.0	199,789.8	1,583.7
CAL	200,141.5	12.3	56.5	200,282.5	0.0	199,818.1	3,875.9
RW	200,116.5	-0.1	34.5	200,282.5	0.0	199,790.1	2,255.3
AC	200,117.1	0.1	26.2	200,282.5	0.0	199,789.8	1,583.7
AR	200,117.3	0.3	23.8	200,282.5	0.0	199,789.2	1,180.7
$\widetilde{AR}$	200,113.1	-1.9	23.8	200,282.5	0.0	199,785.4	1,179.5
Case 3		$\mu(\hat{t}_{z_1})$=199755.597			$sd(\hat{t}_{z_1})$=2193.298		
HT	200,124.8	4.0	34.9	200,329.9	2,204.9	199,755.6	2,193.3
GREGx	200,145.3	14.2	32.1	200,282.5	0.0	199,786.9	1,684.7
GREGxz	200,112.0	-2.4	40.1	200,282.5	0.0	199,755.6	2,193.3
CAL	200,110.9	-3.0	57.4	200,282.5	0.0	199,747.3	3,902.6
RW	200,110.9	-3.0	57.4	200,282.5	0.0	199,747.3	3,902.6
AC	200,119.8	1.5	36.7	200,282.5	0.0	199,755.6	2,193.3
AR	200,112.0	-2.4	40.1	200,282.5	0.0	199,755.6	2,193.3
$\widetilde{AR}$	200,112.0	-2.4	40.1	200,282.5	0.0	199,755.6	2,193.3
Case 4		$\mu(\hat{t}_{z_1})$=199803.988			$sd(\hat{t}_{z_1})$=2208.012		
HT	200,143.4	13.3	35.6	200,201.9	2,129.1	199,850.0	2,223.1
GREGx	200,115.9	-0.4	33.2	200,282.5	0.0	199,798.5	1,764.1
GREGxz	200,115.9	-0.4	40.0	200,282.5	0.0	199,804.0	2,208.0
CAL	200,250.2	66.7	56.3	200,282.5	0.0	199,961.2	3,828.0
RW	200,157.5	20.3	39.8	200,282.5	0.0	199,854.3	2,609.8
AC	200,114.6	-1.1	34.7	200,282.5	0.0	199,804.0	2,208.0
AR	200,119.2	1.2	27.5	200,282.5	0.0	199,803.4	1,386.5
$\widetilde{AR}$	200,114.6	-1.1	27.4	200,282.5	0.0	199,799.4	1,382.9

SCENARIO 21

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.721	-0.975	-0.679	0.129	0.909	0.023	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,797.0	1.4	34.8	200,002.7	2,140.6	200,010.3	2,189.5
GREGx	199,802.7	4.2	24.5	200,013.7	0.0	200,005.1	1,641.2
GREGxz	199,768.5	-12.9	7.1	200,013.7	0.0	200,042.7	0.0
CAL	199,813.5	9.7	26.1	200,013.7	0.0	200,056.3	3,877.3
RW	199,799.7	2.7	26.1	200,013.7	0.0	200,042.7	0.0
AC	199,799.7	2.7	26.1	200,013.7	0.0	200,042.7	0.0
AR	199,768.5	-12.9	7.1	200,013.7	0.0	200,042.7	0.0
$\widetilde{AR}$	199,768.5	-12.9	7.1	200,013.7	0.0	200,042.7	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200073.906			$sd(\hat{t}_{z_1})$=1435.638		
HT	199,786.8	-3.7	34.2	199,997.1	2,168.7	200,067.2	2,136.4
GREGx	199,796.3	1.0	24.0	200,013.7	0.0	200,057.8	1,595.6
GREGxz	199,781.8	-6.2	21.5	200,013.7	0.0	200,073.9	1,435.6
CAL	199,809.4	7.6	25.8	200,013.7	0.0	200,119.0	3,850.5
RW	199,804.2	5.0	25.8	200,013.7	0.0	200,119.3	2,094.0
AC	199,803.0	4.4	25.9	200,013.7	0.0	200,073.9	1,435.6
AR	199,788.0	-3.1	16.8	200,013.7	0.0	200,067.1	1,071.2
$\widetilde{AR}$	199,787.0	-3.6	16.8	200,013.7	0.0	200,068.3	1,068.7
Case 3		$\mu(\hat{t}_{z_1})$=200168.12			$sd(\hat{t}_{z_1})$=2185.373		
HT	199,683.5	-55.5	34.6	200,051.4	2,184.8	200,168.1	2,185.4
GREGx	199,655.4	-69.5	23.9	200,013.7	0.0	200,195.6	1,598.9
GREGxz	199,678.5	-57.9	32.0	200,013.7	0.0	200,168.1	2,185.4
CAL	199,652.1	-71.1	25.6	200,013.7	0.0	200,169.8	3,931.6
RW	199,652.1	-71.1	25.6	200,013.7	0.0	200,169.8	3,931.6
AC	199,648.8	-72.8	25.6	200,013.7	0.0	200,168.1	2,185.4
AR	199,678.5	-57.9	32.0	200,013.7	0.0	200,168.1	2,185.4
$\widetilde{AR}$	199,678.5	-57.9	32.0	200,013.7	0.0	200,168.1	2,185.4
Case 4		$\mu(\hat{t}_{z_1})$=200044.962			$sd(\hat{t}_{z_1})$=2280.853		
HT	199,799.9	2.8	35.5	199,949.1	2,165.2	200,028.5	2,227.7
GREGx	199,844.7	25.3	24.6	200,013.7	0.0	199,986.8	1,638.2
GREGxz	199,791.9	-1.2	33.2	200,013.7	0.0	200,045.0	2,280.9
CAL	199,869.3	37.6	25.9	200,013.7	0.0	200,127.4	3,952.2
RW	199,864.8	35.3	26.0	200,013.7	0.0	200,075.4	2,811.2
AC	199,859.7	32.8	26.1	200,013.7	0.0	200,045.0	2,280.9
AR	199,824.0	14.9	19.8	200,013.7	0.0	200,009.8	1,285.5
$\widetilde{AR}$	199,825.0	15.4	19.7	200,013.7	0.0	200,008.6	1,280.2

SCENARIO 22

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.75	-0.844	0.963	0.185	0.463	0.864	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,916.3	-11.3	34.0	199,970.4	2,316.0	200,068.3	2,250.6
GREGx	199,918.7	-10.1	22.6	199,964.9	0.0	200,064.1	588.6
GREGxz	199,940.6	0.9	16.4	199,964.9	0.0	200,052.6	0.0
CAL	199,953.0	7.0	64.7	199,964.9	0.0	200,064.4	602.9
RW	199,933.0	-2.9	65.0	199,964.9	0.0	200,052.6	0.0
AC	199,933.0	-2.9	65.0	199,964.9	0.0	200,052.6	0.0
AR	199,940.6	0.9	16.4	199,964.9	0.0	200,052.6	0.0
$\widetilde{AR}$	199,940.6	0.9	16.4	199,964.9	0.0	200,052.6	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200167.464			$sd(\hat{t}_{z_1})$=1577.625		
HT	199,875.2	-31.9	37.1	199,972.9	2,245.9	200,091.4	2,271.1
GREGx	199,882.3	-28.3	24.1	199,964.9	0.0	200,082.5	596.1
GREGxz	199,736.0	-101.5	45.3	199,964.9	0.0	200,167.5	1,577.6
CAL	199,909.7	-14.6	66.8	199,964.9	0.0	200,083.9	596.7
RW	199,904.3	-17.3	66.3	199,964.9	0.0	200,061.6	342.7
AC	199,817.0	-61.0	70.4	199,964.9	0.0	200,167.5	1,577.6
AR	199,865.0	-36.9	23.3	199,964.9	0.0	200,092.6	559.0
$\widetilde{AR}$	199,864.6	-37.2	23.3	199,964.9	0.0	200,092.8	559.1
Case 3		$\mu(\hat{t}_{z_1})$=200113.05			$sd(\hat{t}_{z_1})$=2153.458		
HT	199,940.5	0.8	34.5	200,034.9	2,174.3	200,113.1	2,153.5
GREGx	199,998.2	29.7	23.0	199,964.9	0.0	200,044.0	577.2
GREGxz	199,884.1	-27.4	60.2	199,964.9	0.0	200,113.1	2,153.5
CAL	199,912.1	-13.4	63.0	199,964.9	0.0	200,044.1	583.4
RW	199,912.1	-13.4	63.0	199,964.9	0.0	200,044.1	583.4
AC	199,775.3	-81.8	90.9	199,964.9	0.0	200,113.1	2,153.5
AR	199,884.1	-27.4	60.2	199,964.9	0.0	200,113.1	2,153.5
$\widetilde{AR}$	199,884.1	-27.4	60.2	199,964.9	0.0	200,113.1	2,153.5
Case 4		$\mu(\hat{t}_{z_1})$=200169.796			$sd(\hat{t}_{z_1})$=2146.158		
HT	199,926.9	-6.0	33.9	200,003.6	2,138.9	200,086.7	2,136.0
GREGx	199,956.6	8.9	23.5	199,964.9	0.0	200,049.6	594.6
GREGxz	199,757.8	-90.6	59.7	199,964.9	0.0	200,169.8	2,146.2
CAL	199,926.3	-6.3	61.5	199,964.9	0.0	200,049.0	599.2
RW	199,896.9	-21.0	61.5	199,964.9	0.0	200,047.0	418.9
AC	199,809.7	-64.6	68.9	199,964.9	0.0	200,169.8	2,146.2
AR	199,944.0	2.6	23.0	199,964.9	0.0	200,057.2	572.8
$\widetilde{AR}$	199,940.5	0.8	23.0	199,964.9	0.0	200,059.3	571.3

SCENARIO 23

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.2	0.696	-0.631	0.183	0.561	0.488	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,691.0	15.8	35.9	200,145.7	2,158.5	199,672.9	2,239.2
GREGx	199,679.7	10.1	34.9	200,198.6	0.0	199,638.2	1,741.4
GREGxz	199,715.5	28.0	22.5	200,198.6	0.0	199,673.3	0.0
CAL	199,769.0	54.8	54.3	200,198.6	0.0	199,759.0	3,892.8
RW	199,703.7	22.1	23.0	200,198.6	0.0	199,673.3	0.0
AC	199,703.7	22.1	23.0	200,198.6	0.0	199,673.3	0.0
AR	199,715.5	28.0	22.5	200,198.6	0.0	199,673.3	0.0
$\widetilde{AR}$	199,715.5	28.0	22.5	200,198.6	0.0	199,673.3	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199620.931			$sd(\hat{t}_{z_1})$=1442.996		
HT	199,745.1	42.9	34.2	200,243.8	2,191.1	199,696.3	2,251.3
GREGx	199,755.4	48.0	33.2	200,198.6	0.0	199,728.7	1,717.4
GREGxz	199,654.9	-2.3	30.0	200,198.6	0.0	199,620.9	1,443.0
CAL	199,728.9	34.7	53.5	200,198.6	0.0	199,689.9	3,948.5
RW	199,654.0	-2.7	34.7	200,198.6	0.0	199,593.0	2,200.0
AC	199,675.9	8.2	27.8	200,198.6	0.0	199,620.9	1,443.0
AR	199,699.9	20.2	26.8	200,198.6	0.0	199,669.0	1,111.3
$\widetilde{AR}$	199,703.1	21.8	26.7	200,198.6	0.0	199,672.5	1,108.9
Case 3		$\mu(\hat{t}_{z_1})$=199662.349			$sd(\hat{t}_{z_1})$=2265.349		
HT	199,625.5	-17.0	35.1	200,160.0	2,314.8	199,662.3	2,265.3
GREGx	199,617.6	-21.0	34.4	200,198.6	0.0	199,636.4	1,716.7
GREGxz	199,641.6	-9.0	40.7	200,198.6	0.0	199,662.3	2,265.3
CAL	199,692.3	16.4	54.8	200,198.6	0.0	199,739.6	4,081.7
RW	199,692.3	16.4	54.8	200,198.6	0.0	199,739.6	4,081.7
AC	199,635.2	-12.2	37.7	200,198.6	0.0	199,662.3	2,265.3
AR	199,641.6	-9.0	40.7	200,198.6	0.0	199,662.3	2,265.3
$\widetilde{AR}$	199,641.6	-9.0	40.7	200,198.6	0.0	199,662.3	2,265.3
Case 4		$\mu(\hat{t}_{z_1})$=199617.233			$sd(\hat{t}_{z_1})$=2166.74		
HT	199,588.1	-35.8	34.7	200,230.5	2,155.4	199,614.3	2,185.6
GREGx	199,586.4	-36.6	33.7	200,198.6	0.0	199,630.8	1,701.7
GREGxz	199,578.6	-40.5	39.3	200,198.6	0.0	199,617.2	2,166.7
CAL	199,583.6	-38.0	53.4	200,198.6	0.0	199,618.6	3,838.9
RW	199,509.2	-75.3	40.0	200,198.6	0.0	199,524.7	2,631.6
AC	199,578.6	-40.5	34.9	200,198.6	0.0	199,617.2	2,166.7
AR	199,582.9	-38.4	29.7	200,198.6	0.0	199,625.3	1,321.8
$\widetilde{AR}$	199,580.6	-39.5	29.7	200,198.6	0.0	199,622.7	1,320.8

SCENARIO 24

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.496	-0.822	0.554	0.008	0.572	0.087	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,201.4	11.4	34.9	200,231.5	2,366.4	199,777.7	2,272.3
GREGx	200,204.4	12.9	30.2	200,237.5	0.0	199,779.3	1,868.3
GREGxz	200,197.6	9.5	20.0	200,237.5	0.0	199,791.7	0.0
CAL	200,245.7	33.5	61.6	200,237.5	0.0	199,795.3	2,133.0
RW	200,219.3	20.3	61.4	200,237.5	0.0	199,791.7	0.0
AC	200,219.3	20.3	61.4	200,237.5	0.0	199,791.7	0.0
AR	200,197.6	9.5	20.0	200,237.5	0.0	199,791.7	0.0
$\widetilde{AR}$	200,197.6	9.5	20.0	200,237.5	0.0	199,791.7	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199754.155			$sd(\hat{t}_{z_1})$=1542.25		
HT	200,235.5	28.4	35.3	200,179.9	2,297.7	199,768.9	2,216.7
GREGx	200,206.9	14.1	31.0	200,237.5	0.0	199,807.7	1,847.0
GREGxz	200,248.7	35.0	27.5	200,237.5	0.0	199,754.2	1,542.3
CAL	200,327.0	74.1	60.6	200,237.5	0.0	199,837.1	2,102.7
RW	200,309.3	65.3	60.5	200,237.5	0.0	199,799.4	1,246.6
AC	200,298.8	60.0	60.7	200,237.5	0.0	199,754.2	1,542.3
AR	200,233.2	27.2	24.9	200,237.5	0.0	199,774.0	1,184.8
$\widetilde{AR}$	200,232.2	26.8	25.0	200,237.5	0.0	199,775.2	1,184.0
Case 3		$\mu(\hat{t}_{z_1})$=199787.236			$sd(\hat{t}_{z_1})$=2267.624		
HT	200,222.3	21.8	35.8	200,336.0	2,198.6	199,787.2	2,267.6
GREGx	200,267.1	44.2	31.5	200,237.5	0.0	199,734.1	1,893.2
GREGxz	200,224.1	22.7	35.0	200,237.5	0.0	199,787.2	2,267.6
CAL	200,160.9	-8.9	59.6	200,237.5	0.0	199,699.6	2,095.7
RW	200,160.9	-8.9	59.6	200,237.5	0.0	199,699.6	2,095.7
AC	200,137.9	-20.4	57.2	200,237.5	0.0	199,787.2	2,267.6
AR	200,224.1	22.7	35.0	200,237.5	0.0	199,787.2	2,267.6
$\widetilde{AR}$	200,224.1	22.7	35.0	200,237.5	0.0	199,787.2	2,267.6
Case 4		$\mu(\hat{t}_{z_1})$=199742.349			$sd(\hat{t}_{z_1})$=2248.312		
HT	200,185.5	3.4	35.6	200,370.3	2,165.4	199,821.0	2,271.1
GREGx	200,253.4	37.3	31.2	200,237.5	0.0	199,754.2	1,946.8
GREGxz	200,257.8	39.5	34.5	200,237.5	0.0	199,742.3	2,248.3
CAL	200,090.8	-43.9	59.2	200,237.5	0.0	199,700.0	2,158.8
RW	200,079.7	-49.4	59.2	200,237.5	0.0	199,778.4	1,472.0
AC	200,056.8	-60.9	59.7	200,237.5	0.0	199,742.3	2,248.3
AR	200,253.1	37.2	27.1	200,237.5	0.0	199,751.7	1,476.1
$\widetilde{AR}$	200,255.5	38.4	27.1	200,237.5	0.0	199,748.8	1,474.3

SCENARIO 25

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	-0.467	-0.472	0.918	0.009	0.016	0.801	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,398.0	41.5	35.9	199,509.4	2,284.9	199,527.7	2,276.4
GREGx	200,344.3	14.7	31.5	199,627.1	0.0	199,635.8	858.1
GREGxz	200,331.7	8.5	31.4	199,627.1	0.0	199,682.5	0.0
CAL	200,548.2	116.5	61.3	199,627.1	0.0	199,647.1	875.5
RW	200,531.5	108.2	60.9	199,627.1	0.0	199,682.5	0.0
AC	200,531.5	108.2	60.9	199,627.1	0.0	199,682.5	0.0
AR	200,331.7	8.5	31.4	199,627.1	0.0	199,682.5	0.0
$\widetilde{AR}$	200,331.7	8.5	31.4	199,627.1	0.0	199,682.5	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199626.815			$sd(\hat{t}_{z_1})$=1513.756		
HT	200,221.0	-46.8	32.3	199,624.9	2,067.2	199,686.2	2,066.5
GREGx	200,227.6	-43.5	29.4	199,627.1	0.0	199,689.7	883.1
GREGxz	200,231.6	-41.5	30.5	199,627.1	0.0	199,626.8	1,513.8
CAL	200,251.0	-31.9	53.7	199,627.1	0.0	199,690.3	906.1
RW	200,216.8	-48.9	53.2	199,627.1	0.0	199,664.6	482.0
AC	200,176.7	-68.9	55.3	199,627.1	0.0	199,626.8	1,513.8
AR	200,229.4	-42.6	29.5	199,627.1	0.0	199,671.9	757.9
$\widetilde{AR}$	200,229.2	-42.7	29.5	199,627.1	0.0	199,671.9	756.5
Case 3		$\mu(\hat{t}_{z_1})$=199836.132			$sd(\hat{t}_{z_1})$=2211.681		
HT	200,204.8	-54.9	35.2	199,757.8	2,233.8	199,836.1	2,211.7
GREGx	200,260.4	-27.1	31.0	199,627.1	0.0	199,715.9	858.7
GREGxz	200,226.4	-44.1	32.9	199,627.1	0.0	199,836.1	2,211.7
CAL	200,112.5	-101.0	59.5	199,627.1	0.0	199,707.6	880.9
RW	200,112.5	-101.0	59.5	199,627.1	0.0	199,707.6	880.9
AC	200,134.9	-89.8	47.0	199,627.1	0.0	199,836.1	2,211.7
AR	200,226.4	-44.1	32.9	199,627.1	0.0	199,836.1	2,211.7
$\widetilde{AR}$	200,226.4	-44.1	32.9	199,627.1	0.0	199,836.1	2,211.7
Case 4		$\mu(\hat{t}_{z_1})$=199709.227			$sd(\hat{t}_{z_1})$=2215.178		
HT	200,284.2	-15.3	34.9	199,735.5	2,219.3	199,822.9	2,228.7
GREGx	200,330.1	7.6	30.3	199,627.1	0.0	199,720.8	842.7
GREGxz	200,325.5	5.4	32.2	199,627.1	0.0	199,709.2	2,215.2
CAL	200,213.6	-50.5	59.6	199,627.1	0.0	199,716.2	853.1
RW	200,187.7	-63.5	59.8	199,627.1	0.0	199,719.5	616.3
AC	200,172.7	-70.9	64.2	199,627.1	0.0	199,709.2	2,215.2
AR	200,329.0	7.1	30.3	199,627.1	0.0	199,720.2	783.4
$\widetilde{AR}$	200,328.9	7.0	30.3	199,627.1	0.0	199,720.6	782.9

SCENARIO 26

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.547	-0.437	-0.724	0.138	0.005	0.415	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	199,691.6	-37.3	35.6	199,612.0	2,288.2	200,288.6	2,227.6
GREGx	199,711.2	-27.5	29.8	199,658.0	0.0	200,252.8	1,555.1
GREGxz	199,715.9	-25.1	29.8	199,658.0	0.0	200,225.8	0.0
CAL	199,748.0	-9.1	33.8	199,658.0	0.0	200,373.3	4,108.4
RW	199,709.5	-28.3	30.7	199,658.0	0.0	200,225.8	0.0
AC	199,709.5	-28.3	30.7	199,658.0	0.0	200,225.8	0.0
AR	199,715.9	-25.1	29.8	199,658.0	0.0	200,225.8	0.0
$\widetilde{AR}$	199,715.9	-25.1	29.8	199,658.0	0.0	200,225.8	0.0
Case 2		$\mu(\hat{t}_{z_1})$=200314.822			$sd(\hat{t}_{z_1})$=1455.593		
HT	199,752.0	-7.1	35.3	199,762.3	2,214.8	200,224.0	2,231.7
GREGx	199,688.9	-38.6	30.3	199,658.0	0.0	200,302.0	1,520.1
GREGxz	199,686.1	-40.1	30.5	199,658.0	0.0	200,314.8	1,455.6
CAL	199,659.6	-53.3	34.3	199,658.0	0.0	200,163.6	4,054.4
RW	199,691.8	-37.2	32.7	199,658.0	0.0	200,336.5	2,201.1
AC	199,683.4	-41.4	32.1	199,658.0	0.0	200,314.8	1,455.6
AR	199,687.2	-39.5	30.3	199,658.0	0.0	200,312.1	1,030.7
$\widetilde{AR}$	199,687.1	-39.5	30.3	199,658.0	0.0	200,311.9	1,031.4
Case 3		$\mu(\hat{t}_{z_1})$=200203.76			$sd(\hat{t}_{z_1})$=2177.999		
HT	199,698.5	-33.8	34.4	199,686.6	2,199.3	200,203.8	2,178.0
GREGx	199,680.4	-42.9	28.7	199,658.0	0.0	200,224.5	1,521.6
GREGxz	199,685.6	-40.3	29.0	199,658.0	0.0	200,203.8	2,178.0
CAL	199,680.5	-42.9	32.7	199,658.0	0.0	200,214.6	3,968.1
RW	199,680.5	-42.9	32.7	199,658.0	0.0	200,214.6	3,968.1
AC	199,674.6	-45.8	29.9	199,658.0	0.0	200,203.8	2,178.0
AR	199,685.6	-40.3	29.0	199,658.0	0.0	200,203.8	2,178.0
$\widetilde{AR}$	199,685.6	-40.3	29.0	199,658.0	0.0	200,203.8	2,178.0
Case 4		$\mu(\hat{t}_{z_1})$=200220.142			$sd(\hat{t}_{z_1})$=2207.41		
HT	199,848.2	41.1	35.3	199,651.3	2,238.4	200,260.1	2,205.8
GREGx	199,849.4	41.7	29.5	199,658.0	0.0	200,255.5	1,513.6
GREGxz	199,850.6	42.3	29.6	199,658.0	0.0	200,220.1	2,207.4
CAL	199,865.1	49.6	33.2	199,658.0	0.0	200,306.2	4,053.7
RW	199,849.9	42.0	32.1	199,658.0	0.0	200,249.0	2,803.0
AC	199,842.0	38.0	31.8	199,658.0	0.0	200,220.1	2,207.4
AR	199,849.7	41.8	29.4	199,658.0	0.0	200,245.0	1,224.9
$\widetilde{AR}$	199,849.9	41.9	29.4	199,658.0	0.0	200,243.6	1,222.4

SCENARIO 27

	$r(yx)$	$r(yz)$	$r(xz)$	$r(yx\|z)^2$	$r(yz\|x)^2$	$r(xz\|y)^2$	
	0.396	0.178	-0.044	0.168	0.045	0.016	
Stimatore	$\mu(\hat{t}_y)$	Rbias%	RrMSE%	$\mu(\hat{t}_x)$	$sd(\hat{t}_x)$	$\mu(\hat{t}_z)$	$sd(\hat{t}_z)$
Case 1							
HT	200,206.9	-7.7	36.2	199,929.2	2,281.8	199,755.0	2,228.3
GREGx	200,209.9	-6.2	33.4	199,954.2	0.0	199,747.0	2,237.3
GREGxz	200,231.7	4.7	32.7	199,954.2	0.0	199,832.5	0.0
CAL	200,246.0	11.9	39.4	199,954.2	0.0	199,803.5	3,156.0
RW	200,254.8	16.2	34.1	199,954.2	0.0	199,832.5	0.0
AC	200,254.8	16.2	34.1	199,954.2	0.0	199,832.5	0.0
AR	200,231.7	4.7	32.7	199,954.2	0.0	199,832.5	0.0
$\widetilde{AR}$	200,231.7	4.7	32.7	199,954.2	0.0	199,832.5	0.0
Case 2		$\mu(\hat{t}_{z_1})$=199863.124			$sd(\hat{t}_{z_1})$=1527.921		
HT	200,200.5	-10.9	35.1	199,991.0	2,194.9	199,785.0	2,262.4
GREGx	200,192.9	-14.6	32.4	199,954.2	0.0	199,780.4	2,265.5
GREGxz	200,208.5	-6.8	31.9	199,954.2	0.0	199,863.1	1,527.9
CAL	200,177.8	-22.2	38.5	199,954.2	0.0	199,772.0	3,190.9
RW	200,211.3	-5.5	35.1	199,954.2	0.0	199,861.3	1,851.6
AC	200,206.8	-7.7	34.1	199,954.2	0.0	199,863.1	1,527.9
AR	200,203.5	-9.3	31.8	199,954.2	0.0	199,836.3	1,298.6
$\widetilde{AR}$	200,203.4	-9.4	31.8	199,954.2	0.0	199,835.9	1,295.7
Case 3		$\mu(\hat{t}_{z_1})$=199866.334			$sd(\hat{t}_{z_1})$=2212.794		
HT	200,182.6	-19.8	35.5	199,980.3	2,182.6	199,866.3	2,212.8
GREGx	200,174.6	-23.8	32.6	199,954.2	0.0	199,854.5	2,216.2
GREGxz	200,176.8	-22.7	32.6	199,954.2	0.0	199,866.3	2,212.8
CAL	200,169.8	-26.2	38.2	199,954.2	0.0	199,864.7	3,220.6
RW	200,169.8	-26.2	38.2	199,954.2	0.0	199,864.7	3,220.6
AC	200,169.5	-26.3	33.3	199,954.2	0.0	199,866.3	2,212.8
AR	200,176.8	-22.7	32.6	199,954.2	0.0	199,866.3	2,212.8
$\widetilde{AR}$	200,176.8	-22.7	32.6	199,954.2	0.0	199,866.3	2,212.8
Case 4		$\mu(\hat{t}_{z_1})$=199919.045			$sd(\hat{t}_{z_1})$=2259.161		
HT	200,162.3	-29.9	36.7	199,925.0	2,300.1	199,765.7	2,152.7
GREGx	200,176.2	-23.0	33.2	199,954.2	0.0	199,759.2	2,158.3
GREGxz	200,206.6	-7.8	33.4	199,954.2	0.0	199,919.0	2,259.2
CAL	200,205.0	-8.6	38.8	199,954.2	0.0	199,818.8	3,138.5
RW	200,208.1	-7.0	36.3	199,954.2	0.0	199,837.4	2,249.9
AC	200,235.6	6.7	36.8	199,954.2	0.0	199,919.0	2,259.2
AR	200,192.2	-15.0	32.9	199,954.2	0.0	199,841.7	1,561.5
$\widetilde{AR}$	200,191.3	-15.4	32.9	199,954.2	0.0	199,838.7	1,558.9

Bibliografia

Ardilly P.; Tillé Y. (2006). *Sampling methods: excercises and solutions*. Springer, New York.

Ballin M.; Falorsi P. D.; Russo A. (2000). Condizioni di coerenza e metodi di stima per le indagini campionarie sulle imprese. *Rivista di Statistica Ufficiale*, **2**, 31–52.

Bardsley P.; Chambers R. L. (1984). Multipurpose estimations from unbalanced samples. *J. Roy. Stat. Soc. C-App.*, **33**(3), 290–299.

Berger Y. G.; Muñoz J. F.; Rancourt E. (2009). Variance estimation of survey estimates calibrated on estimated control totals - An application to the extended regression estimator and the regression composite estimator. *Comput. Stat. Data An.*, **53**, 2596–2604.

Bethlehem J. G.; Keller J. W. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, **3**(2), 141–153.

Cassel C.; Särndal C.-E.; Wretman J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, **63**(3), 615–620.

Cassel C.; Särndal C.-E.; Wretman J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley & Sons Ltd., New York.

Cassel C. M.; Särndal C.-E.; Wretman J. H. (1979). Prediction theory on finite population when model-based and design-based principle are combined. *Scand. J. Stat.*, **6**(3), 97–106.

Ceccarelli C.; Guandalini A. (2013). Increasing the accuracy of it-silc estimates through the use of auxiliary information from labour force survey. *Accepted in Statistica Applicata - Italian J. Appl. Stat.*

Ceccarelli C.; Giorgi G. M.; Guandalini A. (2010). Lo stimatore di ponderazione vincolata in presenza di informazioni ausiliarie campionarie. Relazione tecnica, Uni-

versità La Sapienza, Dipartimento di Statistica, Probabilità e Scienze Applicate, Roma, Italia.

Ceccarelli C.; Giorgi G. M.; Guandalini A. (2011). Varianza dello stimatore calibrato in presenza di informazioni ausiliarie campionarie. *Riv. Ita. Econom. Demogr. e Stat.*, **65**(1), 53–60.

Ceccarelli C.; Giorgi G. M.; Guandalini A. (2013). Variance of calibrated estimator with dependent estimated control totals. *Submitted to Journal of Official Statistics*.

Chambers R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, **12**(1), 3–32.

Cicchitelli G.; Herzel A.; G.E. M. (1992). *Il campionamento statistico*. Il Mulino, Bologna.

Cochran W. (1977). *Sampling Teechniques*. Wiley & Sons Ltd., New York, 3rd edizione.

Conti P.; Marella D. (2011). *Campinamento da popolazioni finite: teoria e tecnica*. Springer, Milano.

Dever J. A.; Valliant R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, **36**(1), 45–56.

Deville J. C. (1999). Simoultaneus calibration of several surveys. In *Proceedings of Statistics Canada Symposium 99 - Combining data from different sources*, pp. 207–212, Ottawa, Canada. Statistics Canada.

Deville J.-C. (2000). Generalized calibration and application to weighting for non-response In *Compstat - Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*. A cura di Bethlem J., van der Heijden P., pp. 65–76, New York. Springer.

Deville J. C.; Särndal C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, **87**(418), 376–382.

Deville J. C.; Tillé Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, **91**(4), 893–912.

Deville J. C.; Särndal C.-E.; Sautory O. (1993). Generalized raking procedure in survey sampling. *J. Am. Stat. Assoc.*, **88**(423), 1013–1020.

Estevao V.; Särndal C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, **18**(2), 233–255.

Estevao V.; Hidiroglou M.; Särndal C.-E. (1995). Methodological principles for a generalized estimation system at statistics canada. *Journal of Official Statistics*, **11**(2), 181–204.

Fuller W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Fuller W. A. (2002). Regression estimation for survey samples. *Survey Methodology*, **28**(1), 5–23.

Fuller W. A.; Rao J. (2001). A regression composite estimator with application to the canadian labour force survey. *Survey Methodology*, **27**(1), 45–51.

Gini C.; Galvani L. (1929). Di un'applicazione del metodo rappresentativo all'ultimo censimento della popolazione. *Annali di Statistica*, **Serie VI**(4), 1–107.

Goga C. (2008). *Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliarie par des modèles nonparamétriques*. Ph.D dissertation, Haute Bretagna, France.

Goga C.; Shezad M. (2010). Overview of ridge regression estimators in survey sampling. Relazione tecnica, IMB, Université de Bourgogne, Dijon, France.

Guandalini A.; Tillé Y. (2014). The optimal regression estimator calibrated on totals from several surveys. *Submitted to International Statistical Review*.

Guggemos F.; Tillé Y. (2010). Penalized calibration in survey sampling: design-based estimation assisted by mixed models. *J. Stat. Plan. Infer.*, **140**(11), 3199–3212.

Härdle W.; Simar L. (2011). *Applied multivariate statistical analysis*. Springer, Berlin.

Hidiroglou M.; Särndal C.-E. (1998). Use of auxiliary information for two-phase samoling. *Survey Methodology*, **24**, 11–20.

Horvitz D.; Thompson D. (1952). A generalisation of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, **47**(260), 663–685.

Houbiers M. (2004). Toward a social statistical database and unified estimates at statistics netherlands. *Journal od Official Statistics*, **20**(1), 55–75.

Isaki C. T.; Fuller W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.*, **77**(377), 89–96.

Istat (2006). La rilevazione sulle Forze di Lavoro: contenuti, metodologie, organizzazione. *Metodi e Norme*, **32**.

Istat (2008). L'indagine europea sui redditi e le condizioni di vita delle famiglie (It-Silc). *Metodi e Norme*, **32**.

Kim J.; Li J.; Valliant R. (2007). Cell collapsing in poststratification. *Survey Methodology*, **33**(2), 139–150.

Kish L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, **8**(2), 183–200.

Knottnerus P.; van Duin C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, **22**(3), 565–584.

Kott P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**(2), 133–142.

Kroese A.; Ressen R. (1999). Weighting and imputation at statistics netherland. In *Proceedings of the IASS Conference of Small Area Estimation, Riga*, pp. 109–120.

Lemaître G.; Dufour J. (1987). An integrated method for weighting persons and families. *Survey methodology*, **13**(2), 199–207.

Lessler J.; Kalsbeek W. (1992). *Non sampling error in surveys*. Wiley & Sons Ltd., New York.

Luery D. (1986). Weighting survey data under linear constraints on the weights. pp. 325–330. American Statistical Associacion.

Mardia K.; Kent J.; Bibby J. (1979). *Multivariate Analysis*. Academic Press, London.

Merkouris T. (2004). Combining independent regression estimators from multiple surveys. *J. Am. Stat. Assoc.*, **99**(468), 1131–1139.

Merkouris T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *J. R. Statist. Soc. B*, **72**(Part 1), 27–48.

Montanari C. (1987). Post-sampling efficient prediction in large-scale survey. *Int. Stat. Rev.*, **55**(2), 191–202.

Montanari C. (1988). On regression estimation of finite population mean. *Survey Methodology*, **24**(1), 69–77.

Nicolini G.; Marasini D.; Montanari G.; Pratesi M.; Ranalli M.; Rocco E. (2013). *Metodi di stima in presenza di errori non campionari*. Springer-Verlag, Milano.

Qualité L.; Tillé Y. (2008). Variance estimation of changes in repeated surveys and its application to the swiss survey of value added. *Survey Methodology*, **34**(2), 173–181.

Rancourt E. (2001). La régression étendue: un ensemble de pratiques d'estimation qui poussent constamment la théorie. Recuil Enquêtes In *Modéles et Application*. A cura di Droesbeke J. J., Lebart I., pp. 334–343, Dunod. (in French).

Rao J. (1973). On double sampling for stratification and analytical survey. *Biometrika*, **60**(1), 125–133.

Rao J. (1994). Estimating totals and distribution functions usinf auxiliary information at estimation stage. *Journal of Official Statistics*, **10**(2), 153–165.

Rao J. (2003). *Small Area Estimation*. Wiley & Sons Ltd., Hoboken, New Jersey.

Rao J. N. K.; Singh A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Method*, pp. 207–212, Washington, D.C. American Statistical Association.

Renssen R. H.; Nieuwenbroek N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *J. Am. Stat. Assoc.*, **92**(437), 368–374.

Robinson P. M.; Särndal C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya, Series B*, **45**(2), 240–248.

Särndal C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, **67**(3), 639–650.

Särndal C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**(2), 99–119.

Särndal C.-E.; Lunström S. (2005). *Estimation in survey with nonresponse*. Wiley & Sons Ltd., Chichester, England.

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

Seber G. (1977). *Linear Regression Analysis*. Wiley & Sons Ltd., New York.

Singh A. C.; Mohl C. A. (1996). Understanding calibration estimator in survey sampling. *Survey Methodology*, **22**(2), 107–115.

Singh A. C.; Kennedy B.; Wu S. (2001). Regression composite estimation for the canadian labour force survey with rotating panel design. *Survey Methodology*, **27**(1), 33–44.

Stukel D. M.; Hidiroglou M. A.; Särndal C. (1996). Variance estimation for calibration estimator: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, **22**(2), 117–125.

Tam S. (1984). On covariance from overlapping samples. *The American Statistician*, **38**(4), 288–289.

Théberge A. (1999). Calibration and restricted weights. *Survey Methodology*, **26**(1), 99–107.

Théberge A. (2000). Estension of calibration in survey sampling. *J. Am. Stat. Assoc.*, **94**(446), 635–644.

Traat I.; Särndal C.-E. (2011). Domain estimators calibrated on information from other surveys. Acta e commentationes universitatis tatuensis de mathemathica 2, University of Tartu, Tartu, Estonia.

van Duin C.; Snijder V. (2010). Simulation studies of repeated weighting. Relazione tecnica, Statistics Netherland, Voorburg: Netherland.

Vitali O. (1993). *Statistica per le scienze applicate*, volume Volume secondo. Cacucci Editore, Bari.

Wolter K. M. (1979). Composite estimation in finite population. *J. Am. Stat. Assoc.*, **74**(367), 604–613.

Wright R. L. (1983). Finite population sampling with multivariate auxiliary information. *J. Am. Stat. Assoc.*, **78**(384), 879–884.

Zieschang K. D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. In *Proceedings of the survey research methods section*, pp. 64–71, Washington, D.C. American Statistical Association.

Zieschang K. D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *J. Am. Stat. Assoc.*, **85**(412), 886–1001.