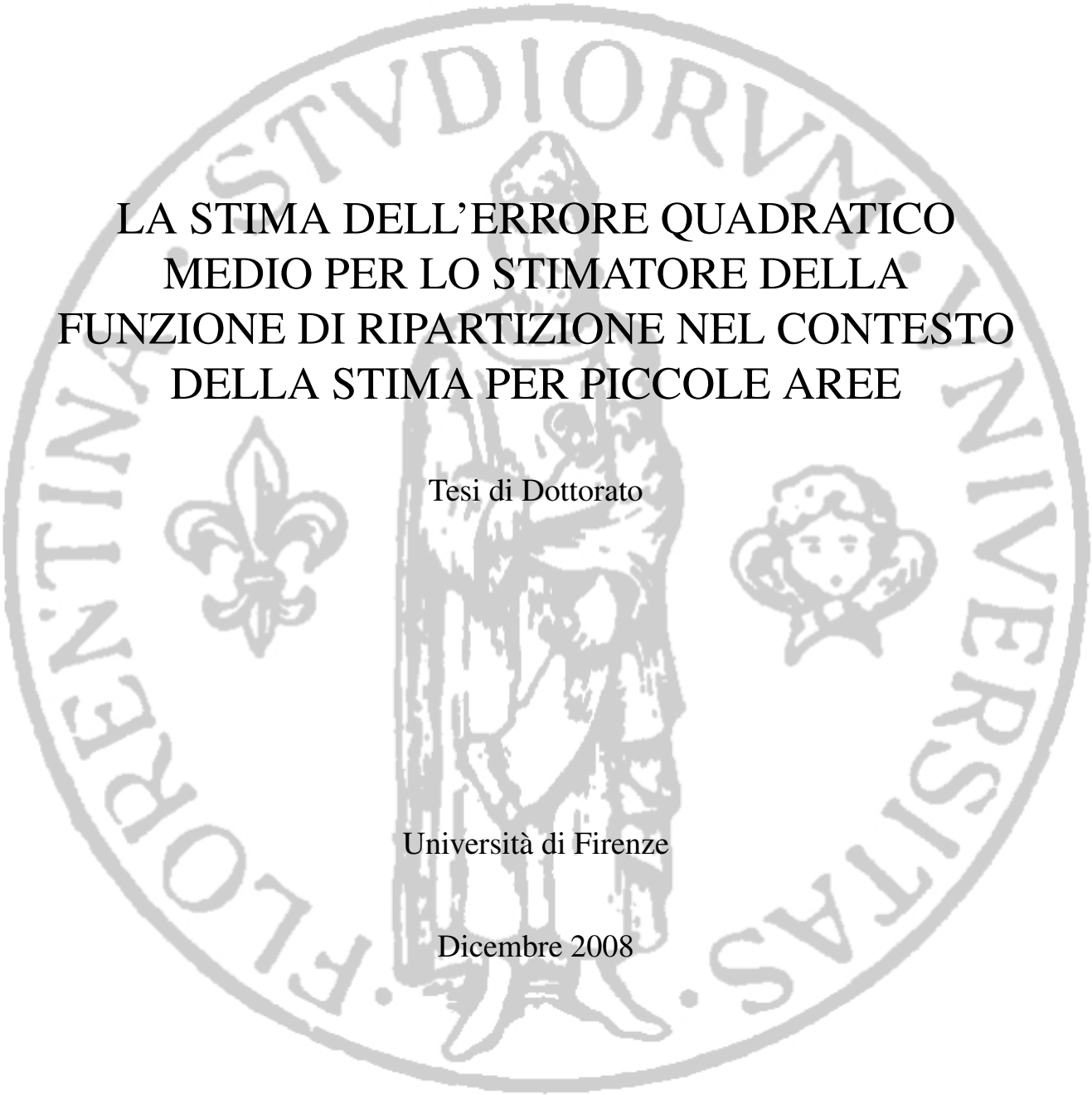
Stefano Marchetti



LA STIMA DELL'ERRORE QUADRATICO MEDIO PER LO STIMATORE DELLA FUNZIONE DI RIPARTIZIONE NEL CONTESTO DELLA STIMA PER PICCOLE AREE

Tesi di Dottorato

Università di Firenze

Dicembre 2008

Università degli Studi di Firenze
Dipartimento di Statistica G. Parenti
Dottorato di Ricerca in Statistica Applicata
XXI ciclo

La stima dell'errore quadratico medio per lo stimatore della funzione di ripartizione nel contesto della stima per piccole aree

Autore
Stefano MARCHETTI

Tutor
Prof.ssa Monica PRATESI

Co-Tutor
Prof. Luigi BIGGERI

Coordinatore
Prof. Guido FERRARI

Settore scientifico-disciplinare: *SECS-S/01 - Statistica*

Alla mia famiglia

Sommario

L'obiettivo di questo lavoro è quello di proporre uno stimatore dell'errore quadratico medio per lo stimatore della funzione di ripartizione proposta da Chambers e Dunstan (1986) nell'ambito della stima per piccole aree.

La crescente domanda da parte di operatori pubblici e privati di informazioni mirate su specifici domini, o aree geografiche, legata al costo elevato delle indagini statistiche e alla frequenza decennale dei censimenti ha aperto un filone di ricerca noto come "stima per piccole aree". Alla richiesta di dati per piccoli domini o aree si affianca la domanda di statistiche più esaustive rispetto agli indicatori tradizionali (come la media), soprattutto per quei fenomeni che presentano dinamiche complesse. In questo lavoro si presentano, unitamente alle tecniche di stima per piccole aree, alcune tecniche di stima che consentono di avere una panoramica completa di una variabile oggetto di interesse, dove per panoramica completa si intende la conoscenza della funzione di ripartizione o di densità della variabile stessa.

Nonostante i metodi di stima della funzione di ripartizione descritti nella tesi siano metodi generici, l'attenzione è stata posta su quei metodi che consentono di ottenere una stima per piccole aree della funzione di ripartizione, per questo si inquadra questo lavoro nell'ambito della stima per piccole aree. Inoltre, lo stimatore della funzione di ripartizione proposto da Chambers e Dunstan (1986) si basa esclusivamente su modello e prevede l'utilizzo di variabili ausiliarie note per tutte le unità della popolazione. Per questi motivi la tesi si colloca nel filone di ricerca noto come metodi di stima per piccole aree basati su modello a livello di unità.

Nel primo capitolo della tesi si introduce il problema della stima per piccole aree, con una panoramica sui metodi di stima attualmente più diffusi.

Nel secondo capitolo si descrivono due approcci alla stima per piccole aree basata su modello: l'approccio basato sul modello lineare ad effetti misti (descritto in modo dettagliato in Rao (2003)) e l'approccio basato sul modello di regressione M-quantile (Chambers e Tzavidis, 2006). Dopo una breve panoramica sui modelli lineari ad effetti misti e su un loro utilizzo nell'ambito della stima per piccole aree, si descrivono quegli strumenti metodologici che saranno utilizzati per la stima della funzione di ripartizione per piccola area e per la stima del suo errore quadratico medio: la *regressione quantilica* (Koenker e Bassett, 1978), gli stimatori robusti *M-estimator* (Huber, 1981) e gli *asymetric least squares* (Newey e Powell, 1987), che sono propedeutici per descrivere il modello di *regressione M-quantile* (Breckling e Chambers, 1988).

Il terzo capitolo tratta dei metodi di stima per la funzione di ripartizione, in particolar modo l'attenzione è posta sullo stimatore della funzione di ripartizione proposto da Chambers e Dunstan (1986) e ad una sua estensione all'ambito della stima per piccole aree. Questo metodo di stima si basa su un modello di superpopolazione che utilizza alcune variabili ausiliarie note, in modo certo e senza errore, per tutte le unità della popolazione. In questo capitolo si presenta anche lo stimatore per la funzione di ripartizione proposto da Rao *e altri* (1990) adattato alla stima per piccole aree.

Nel quarto capitolo si presenta una proposta originale per la stima dell'errore quadratico medio

dello stimatore della funzione di ripartizione proposto da Chambers e Dunstan (1986) nell'ambito della stima per piccola area. Lo stimatore in questione si basa su una procedura bootstrap proposta da Lombardia *e altri* (2003) per la stima dell'errore quadratico medio dello stimatore della funzione di ripartizione proposto da Chambers e Dunstan (1986). Le proprietà dello stimatore dell'errore quadratico medio dello stimatore della funzione di ripartizione sono state verificate empiricamente tramite due simulazioni Monte Carlo: una basata su modello e una basata su disegno. La stima della funzione di ripartizione (Chambers e Dunstan, 1986) e del suo errore quadratico medio (proposto in questo lavoro) è stata applicata ai dati dell'indagine *European Union - Statistics on Income and Living Conditions (EU-SILC)* del 2004. Sono stati stimati per le province della regione Toscana alcuni percentili del reddito equivalente disponibile nell'anno 2003, rilevato nell'indagine EU-SILC del 2004, utilizzando come fonte di dati ausiliaria il *Censimento Famiglie e Abitazioni* del 2001. Utilizzando lo stimatore proposto per la stima dell'errore quadratico medio della funzione di ripartizione si è stimato l'errore quadratico medio dei percentili stimati nelle dieci province della regione Toscana.

Indice

Elenco delle tabelle

Capitolo 1

La Stima per Piccole Aree

1.1 Introduzione

Ai metodi di stima per piccole aree viene attualmente dedicata grande attenzione a livello internazionale vista la crescente domanda da parte di enti pubblici e privati di informazioni statistiche precise e tempestive relative a territori e domini ridotti.

Le fonti statistiche censuarie, come i censimenti della popolazione, dell'agricoltura, dell'industria e del commercio, forniscono dati totalmente esaustivi molto utili agli operatori pubblici e privati per programmare e verificare lo svolgimento delle loro attività, ma allo stesso tempo, i censimenti, hanno periodicità decennale e trattano un numero limitato di aspetti. Anche gli archivi amministrativi forniscono informazione esaustive su tutta la popolazione a cui si riferiscono, ma alcuni problemi di definizione degli ambiti territoriali di riferimento e l'affidabilità stessa dei dati presenti negli archivi amministrativi (in Italia), rendono preferibile l'utilizzo di questi dati come fonte ausiliaria (Chiandotto, 1996).

Per questi motivi le rilevazioni campionarie hanno assunto un ruolo fondamentale nel processo conoscitivo e decisionale per le istituzioni pubbliche e gli enti pubblici e privati. Nelle indagini su vasta scala, però, i dati sono solitamente ottenuti sulla base di un disegno campionario complesso e le dimensioni campionarie adeguate per garantire l'affidabilità delle stime ad un livello prestabilito di ripartizione geografica o dominio. Di conseguenza non è possibile ottenere stime relative a aree geografiche o domini ridotti e non considerati come domini al momento del disegno dell'indagine (per esempio l'indagine promossa dall'Unione Europea sui redditi e sulle condizioni di vita delle famiglie (EU-SILC) è progettata, in Italia, per ottenere stime affidabili a livello regionale, allora le province, i comuni, i lavoratori dipendenti, le donne, e altri sottodomini sono domini non considerati). Utilizzando stimatori basati solo sul disegno campionario su questi domini si potrebbero ottenere stime scarsamente affidabili.

La richiesta da parte degli operatori pubblici e privati di informazioni accurate e tempestive per fini decisionali e operativi non può essere soddisfatta isolatamente dalle rilevazioni censuarie, dagli archivi amministrativi e da quelle campionarie. Per questo motivo sono stati sviluppati metodi di stima che combinano le diverse fonti con l'obiettivo di migliorare la precisione e l'affidabilità delle stime. La soluzione al problema di carenza di informazioni a livello di domini non previsti si collocano in due diversi filoni di ricerca:

1. l'incremento della numerosità del campione fino ad ottenere il livello di precisione desiderato per l'area geografica o il dominio oggetto di interesse. Questo comporta in fase di progettazione dell'indagine la definizione dell'area o dominio oggetto di studio per la quale si vogliono

produrre stime e in alcuni casi un elevato aumento dei costi di rilevazione e dei tempi di analisi dei dati;

2. la formulazione di stimatori che consentono di migliorare l'efficienza delle stime rispetto a quello che si otterrebbe sulla base del disegno di campionamento.

Il 2^o filone di ricerca è quello che ha riscontrato più successo poiché offre almeno due vantaggi molto importanti. Primo, non è necessario conoscere il dominio di interesse in fase di progettazione dell'indagine. Secondo, si riducono i costi di rilevazione, che in certi ambiti di ricerca (come quello socio-economico) sono molto alti.

Prima di presentare una panoramica dei metodi di stima per piccole aree è necessario definire in modo univoco il significato del termine piccola area.

Un primo tentativo di definizione di piccola area fu fatto da Purcell e Kish (1980) che consideravano piccole aree quelle formate da un numero di unità statistiche comprese tra $1/10$ e $1/100$ della popolazione di riferimento. In seguito Brackstone (1987) ha definito piccola area o dominio qualunque area in cui stime accurate non possono essere derivate utilizzando informazioni provenienti da rilevazioni campionarie correnti, ma nuovi metodi di stima sono necessari.

Rao (2003) afferma che il termine piccola area è impiegato per aree geografiche di piccole dimensioni (come province, comuni, sezioni di censimento, etc.) e per descrivere piccoli domini formati da sub-popolazioni con certe caratteristiche di età, sesso e razza all'interno di un'ampia area geografica, come ad esempio una nazione. Seguendo l'indirizzo di Brackstone (1987), Rao (2003) definisce un'area come piccola se la numerosità campionaria specifica di quell'area non consente di ottenere stime dirette di adeguata precisione.

In questo lavoro si adotta la definizione di piccola area proposta da Rao (2003).

1.2 I metodi di stima per piccole aree: una panoramica

A partire dagli anni '80 in poi, nella letteratura specializzata, ci sono state molte proposte di metodi di stima per piccole aree. Una classificazione dei metodi di stima per piccola area presenti in letteratura è stata fatta, in Italia, da Gori e Marchetti (1987). Tale classificazione i basa sulla diversità dei metodi di stima piuttosto che sulla struttura degli stimatori. La classificazione proposta da Gori e Marchetti (1987) è la seguente:

i. metodi campionari, che senza assumere esplicitamente un modello, forniscono direttamente l'espressione di stimatori per il parametro incognito caratteristico di piccola area (metodi demografici, stimatori sintetici, metodi di regressione campionari),

ii. metodi basati su modello, che prevedono l'assunzione esplicita di un modello di superpopolazione (approccio predittivo, metodi per dati qualitativi, modelli bayesiani),

iii. metodi composti che combinano, attraverso medie ponderate, stimatori del tipo **i.** e **ii.** con stimatori diretti ottenuti da disegni campionari.

Tuttavia gli stessi autori (Gori e Marchetti, 1987) riconoscono che la distinzione tra i metodi **i.** e **ii.** non è molto rigorosa e deriva soprattutto da ragioni di ordine storico. Una rassegna approfondita dei metodi di stima per piccole aree dal punto di vista teorico e applicativo è stata proposta da Russo (1996).

Sulla base dei dati utilizzati Chiandotto (1996) propone una classificazione, quasi universalmente riconosciuta, tra stimatori diretti e stimatori indiretti. Si parla di stimatori diretti se si impiegano esclusivamente i valori della variabile oggetto di studio ottenuti sulle sole unità campionarie appartenenti alla piccola area e relativi alla rilevazione corrente. Se, invece, per realizzare stime per piccole aree con un adeguato livello di precisione si "pende forza in prestito" (dall'inglese *borrow strength*) dai valori della variabile di interesse osservata sulle unità campionarie di un'area contenente la piccola area e/o relativi ad altre occasioni d'indagine, oltre a quella corrente, si definiscono gli stimatori indiretti.

Una distinzione funzionale dei metodi di stima per piccola area si basa sul tipo di inferenza. La seguente distinzione dei metodi di stima per piccole aree classificati secondo il tipo di inferenza è stata proposta, tra gli altri, da Salvati (2004):

- *Metodi di stima basati su disegno (o campionari)*: la stima del parametro di interesse di piccola area è ottenuta attraverso l'utilizzo dei metodi campionari classici basati sulla distribuzione di probabilità indotta dal disegno di campionamento. Un aspetto fondamentale di questa impostazione è costituito dal fatto che il parametro, o sue funzioni, è pensato come una costante. Inoltre gli stimatori sono corretti rispetto al disegno di campionamento applicato. Purtroppo la loro variabilità cresce al diminuire della numerosità campionaria e può accadere che nessuna unità campionaria sia presente nella piccola area impedendo di ottenere una stima del parametro di interesse di piccola area.

  Questa classe è composta solo da metodi diretti e ne fanno parte gli stimatori di espansione nell'ambito dei quali il più utilizzato è senz'altro quello di Horvitz e Thompson (1952).

- *Metodi assistiti da modello*: il termine è stato coniato da Särndal (1993) per definire quei metodi per i quali l'inferenza è basata sia sul disegno sia sul modello. L'obiettivo è quello di ottenere, sfruttando l'informazione derivante dal disegno di campionamento, stimatori corretti indipendentemente dalla scelta del modello, che però assume importanza per introdurre le ipotesi fatte dal ricercatore sul legame fra variabili ausiliarie e variabile di interesse.

  Questa classe è formata dallo stimatore diretto di regressione, e da molti stimatori indiretti tra i quali rivestono maggiore importanza gli stimatori sintetici e quelli combinati.

- *Metodi basati su modello*: tale impostazione è nota in letteratura anche come approccio predittivo. L'aspetto saliente è costituito dal fatto che il parametro oggetto di studio, o sue funzioni, non è pensato come una costante, ma è visto, invece, come una variabile casuale. L'approccio prevede l'introduzione di un modello probabilistico di superpopolazione relativo alla distribuzione del fenomeno tra le aree da cui derivare il predittore ottimo e corretto a livello di piccola area.

  Appartengono a questa categoria i modelli di piccola area (*Small Area Methods*) che prevedono la presenza di effetti casuali di area (per esempio *Area Level Random Effects Model* (Fay e Herriot, 1979)) e *Nested Error Unit Level Regression Model* (Battese, 1988)).

1.2.1 I metodi di stima per piccola aree basati su disegno

Generalmente le piccole aree di interesse sono rappresentate da suddivisioni geografico-amministrative che hanno una struttura gerarchica, che in Italia è rappresentata da regioni, province e comuni. L'interesse di molti enti è spesso rivolto alla stima per province (Falorsi *e altri*, 1994), ma non sono rari i casi in cui il dettaglio desiderato sia addirittura sub-comunale (Giommi e Rocco, 2003). E' raro poter definire tutte le aree di interesse in fase di progettazione della rilevazione. Come conseguenza, le

stime per le piccole aree di interesse non sono disponibili oppure, quando ci sono, sono scarsamente affidabili (Singh *e altri*, 1994)

In genere condizionandosi ai dati disponibili è possibile costruire stime basate su disegni sia dirette sia indirette. Lo stimatore diretto del valore di interesse $f(Y)_i$ relativo all'area i-esima è ottenuto sulla base dei valori della variabile di studio osservati con riferimento alle sole unità del campione appartenenti alla piccola area i. Tale stimatore è corretto sulla base del disegno, ma può essere caratterizzato da un'elevata variabilità. Gli stimatori indiretti si avvalgono di valori della variabile di interesse osservati su aree diverse da quella di interesse, oppure rilevati sulla stessa area ma riferiti a tempi diversi.

Gli stimatori diretti più diffusi sono:

- lo stimatore Horvitz-Thompson per area,
- lo stimatore post-stratificato per area,
- lo stimatore rapporto per area.

Per un approfondimento su questi stimatori si consulti Särndal *e altri* (1992).

1.2.2 I metodi di stima per piccola area assistiti da modello

In questo approccio alla stima per piccole aree si sfrutta il legame, espresso da un modello relazionale, tra alcune variabili ausiliarie osservabili e la variabile di studio riferita alla piccola area. Questi metodi di stima sono basati sul disegno di campionamento dal quale si ricavano degli opportuni pesi campionari che garantiscono una correttezza approssimata degli stimatori rispetto al disegno, sia nel caso che il modello relazionale si adatti bene ai dati, sia nel caso contrario. I principali metodi di stima per piccole aree assistiti da modello sono **i.** lo stimatore di regressione per piccola area, **ii.** lo stimatore sintetico e **iii.** lo stimatore combinato.

Lo stimatore di regressione per piccole aree è stato introdotto da Särndal *e altri* (1992) ed altro non è che uno stimatore di regressione basato sul disegno di campionamento applicato alla stima per piccole aree. Il suo utilizzo prevede di conoscere le variabili ausiliarie in modo certo e senza errore per tutte le unità della popolazione obiettivo.

Lo stimatore sintetico è uno stimatore indiretto che, utilizzando la stima diretta della variabile di interesse per una "grande" area[1], deriva stime per piccole aree (che sono sub-aree della grande area) assumendo che le piccole aree siano simili, per certe caratteristiche, alla grande area che le contiene (Gonzalez, 1973). Lo stimatore sintetico contempla sia la possibilità di non utilizzare variabili ausiliarie sia la possibilità di utilizzarle.

Nel caso che non siano disponibili variabili ausiliarie lo stimatore sintetico si basa sull'assunzione che le medie di piccola area sono uguali alla media della grande area che le contiene, tale ipotesi è denominata ipotesi di omogeneità. Quando l'ipotesi di omogeneità non è vera lo stimatore può presentare distorsioni di notevole entità.

Nel caso in cui siano disponibili informazioni ausiliarie a livello di piccola area lo stimatore sintetico si basa sul modello di regressione lineare: la quantità incognita della piccola area i è stimata dal

[1] Per grande area si intende un area geografica o dominio considerato al momento della progettazione dell'indagine e per cui la stima diretta fornisce stime affidabili.

vettore di variabili ausiliarie della piccola area i e da un coefficiente di regressione comune a tutte le piccole aree stimato su tutto il campione.

Lo stimatore sintetico è generalmente distorto soprattutto perché è rara l'ipotesi di omogeneità (Rao, 2003; Gonzalez, 1973); inoltre è difficile ottenere una stima dell'errore quadratico medio perché è difficile stimare la distorsione dello stimatore sintetico.

Possono rientrare nella classe di stimatori sintetici anche i metodi demografici per piccole aree. Questi metodi si basano sull'utilizzo delle fonti censuarie più recenti integrate con dati amministrativi a livello locale. I principali metodi demografici sono: il metodo dei tassi di sopravvivenza (Bogue, 1950), il metodo composito (Bogue e Duncan, 1959), il metodo delle componenti (of the Census, 1966) e i metodi di regressione campionari (Ericksen, 1974)[2].

Lo stimatore combinato si ottiene come media ponderata tra uno stimatore diretto e uno stimatore indiretto. Uno stimatore diretto è ottenuto sulla base dei valori della variabile di studio osservati con riferimento alle sole unità del campione appartenenti alla piccola area i. Tale stimatore è corretto rispetto al disegno di campionamento, ma può essere caratterizzato da una variabilità molto elevata. Uno stimatore indiretto utilizza le variabili ausiliarie per ridurre la variabilità degli stimatori diretti. Le informazioni ausiliarie possono essere di vario tipo: relative alla piccola area, relative alla grande area, disponibili per le unità campionate o disponibili per tutte le unità della popolazione. Il legame tra la variabile ausiliaria e la variabile di studio è formalizzata, in genere, attraverso un modello. Se $f(\hat{Y})_i^D$ indica lo stimatore diretto per l'area i e $f(\hat{Y})_i^I$ indica lo stimatore indiretto per l'area i, allora lo stimatore combinato, $f(\hat{Y})_i^C$, è:

$$f(\hat{Y})_i^C = \gamma_i f(\hat{Y})_i^D + (1-\gamma_i) f(\hat{Y})_i^I,$$

dove γ_i rappresenta un peso opportunamente scelto ($\gamma_i \in [0,1]$). Solitamente il peso γ_i è ottenuto minimizzando l'errore quadratico medio dello stimatore combinato rispetto al peso stesso assumendo che non ci sia correlazione lineare tra lo stimatore diretto e lo stimatore indiretto (Rao, 2003; Shaible, 1978)[3]. Un metodo alternativo per stabilire il peso γ_i è stato proposto da Drew, Singh e Chouchry in Gosh e Rao (1994). In questa classe di stimatori rientrano molti degli stimatori usati nell'ambito della stima per piccole aree, come ad esempio lo stimatore EBLUP[4].

1.2.3 I metodi di stima per piccola area basati su modello

I metodi di stima per piccola are basati su modello si fondano su modelli probabilistici che assumono effetti casuali specifici di area per esprimere la variabilità dei valori della variabile di interesse tra aree, oltre a quella spiegata dalle variabili ausiliarie incluse nel modello (Pfeffermann, 2002). I parametri del modello possono essere stimati con un approccio classico-frequentista oppure bayesiano. Datta e Lahiri (1997), e altri autori, affermano la superiorità dei metodi bayesiani nel problema della stima per piccole aree, ed in particolare dei metodi bayesiani gerarchici che, anche se più impegnativi dal punto di vista algebrico-computazionale, forniscono stime migliori rispetto agli stimatori bayesiani empirici poiché non trattano l'incertezza relativa ai parametri incogniti presenti nel modello (Datta e Gosh, 1991). Rao (2003) sostiene che l'impiego dell'approccio bayesiano non sia conveniente qualora si utilizzi una priori non informativa sui parametri del modello.

I modelli di stima per piccole aree basati su modello sono classificati in due categorie:

[2] Per una trattazione esaustiva si veda Rao (2003).

[3] γ_i è denominato in letteratura fattore di restringimento o *shrinkage factor*.

[4] Predittore empirico lineare e corretto (*Empirical Best Linear Unbiased Predictor*) (Rao, 2003).

- *Modelli a livello di area*: legano la stima diretta di piccola area con i valori di covariate specifiche dell'area stessa. Si utilizzano questi modelli quando non sono disponibili informazioni ausiliarie a livello di unità campionarie,
- *Modelli a livello di unità*: mettono in relazione i valori della variabile di studio ottenuti dalle osservazioni campionarie con i valori delle variabili ausiliarie disponibili a livello di unità campionarie.

Il principale punto di debolezza dei metodi di stima basati su modello è dovuto alla scelta del modello stesso. Infatti il processo di inferenza si basa sulla distribuzione probabilistica imposta dal modello scelto. Se il modello utilizzato non si adatta bene ai dati è possibile ottenere delle stime distorte. D'altronde i metodi di stima basati su modello presentano diversi vantaggi (Salvati, 2004):

- sotto il modello assunto si possono ottenere stimatori ottimi per piccola area,
- ad ogni stima per piccola area si può associare una stima di variabilità (Pfeffermann, 2002),
- il modello scelto può essere validato dai dati campionari,
- si può scegliere il modello che più si adatta al tipo di variabile oggetto di studio,
- il modello permette la stima del parametro di interesse anche per le aree non campionate[5].

I metodi di stima basati su modello sono molto usati nella stima per piccole aree; anche in questo lavoro si affronta la stima per piccole aree della funzione di ripartizione e dei quantili con una metodologia basata su modello. Vista l'importanza che ricoprono i metodi di stima basati su modello se ne darà una descrizione formale, anche se non dettagliata, nel capitolo 2.

[5]Le piccola aree in cui non sono presenti unità campionate.

Capitolo 2

Modelli per la stima per piccole aree

2.1 Introduzione

La stima per piccole aree basata su modello si avvale generalmente di modelli relazionali per predire il valore della variabile di studio per le unità non campionate sfruttandone la relazione con alcune variabili ausiliarie note. Esistono diversi "scenari" possibili per quanto concerne la conoscenza o meno delle variabili ausiliarie. In questo testo si presume che tutte le variabili ausiliarie rilevanti siano osservabili e note senza errore per tutte le unità della popolazione.

I modelli maggiormente impiegati nella stima per piccole aree sono i modelli lineari ad effetti misti, ampiamente utilizzati in questo contesto, ed i modelli di regressione M-quantile, proposti recentemente in letteratura (Chambers e Tzavidis, 2006).

Considerando la diffusione dei modelli lineari ad effetti misti, verrà illustrato, molto brevemente, il modo con cui vengono impiegati nella stima per piccole aree.

Una trattazione più accurata sarà riservata al modello di regressione M-quantile. Per presentare le caratteristiche di tale modello sarà necessario descrivere diverse metodologie statistiche: la *regressione quantilica*, l'*M-estimator* e l'*expectile*. Come si vedrà, l'utilizzo del modello di regressione M-quantile offre notevoli vantaggi rispetto al modello lineare ad effetti misti, per questo si ritiene utile approfondire la ricerca su questo modello.

2.2 Il modello lineare ad effetti misti applicato alla stima per piccole aree

L'obiettivo di questo paragrafo è quello di introdurre le ipotesi di base del modello lineari ad effetti misti e la sua applicazione nell'ambito della stima per piccole aree.

Il modello è generalmente espresso usando la seguente formulazione:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{2.1}$$

dove $\boldsymbol{y}$ è il vettore nx1 della variabile di interesse, $\boldsymbol{\beta}$ è il vettore px1 dei coefficienti di regressione degli effetti fissi, $\boldsymbol{X}$ è la matrice nxp di p variabili ausiliarie, $\boldsymbol{u}$ è il vettore qx1 degli effetti casuali, $\boldsymbol{Z}$ è la matrice nxq delle variabili associate agli effetti casuali ed $\boldsymbol{e}$ è il vettore nx1 degli errori casuali. n è il numero delle osservazioni note per $\boldsymbol{y}$ e $\boldsymbol{X}$. Nella (2.1) si assume che gli effetti casuali e gli errori siano indipendenti e normalmente distribuiti con media zero e matrice di varianza nota, dipendente da alcuni parametri $\boldsymbol{\theta}$, denominati componenti di varianza,

$$V[\boldsymbol{u}] = E[\boldsymbol{u}\boldsymbol{u}'] = \boldsymbol{V}_u$$
$$V[\boldsymbol{e}] = E[\boldsymbol{e}\boldsymbol{e}'] = \boldsymbol{V}_e.$$

Dalla (2.1) si ottiene la matrice di varianza per $\boldsymbol{y}$:

$$V[\boldsymbol{y}] = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}' + \boldsymbol{V}_e.$$

La stima BLUE per $\boldsymbol{\beta}$ si può ricavare sia con il metodo della massima verosimiglianza, sia con il metodo dei minimi quadrati. In entrambi i casi

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.$$

La stima BLUP per $\boldsymbol{u}$ è:

$$\hat{\boldsymbol{u}} = \boldsymbol{V}_u\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

La forma del modello lineare ad effetti misti principalmente usato nella stima per piccole aree è quella nota come ANOVA,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{m}\boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}, \tag{2.2}$$

dove $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{X}$ ed $\boldsymbol{e}$ mantengono lo stesso significato che avevano nel modello (2.1), mentre $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}, \ldots, \boldsymbol{u}_{iq_i})'$ è il vettore dei q_i effetti casuali del fattore i-esimo e $\boldsymbol{Z}_i$ è una matrice di costanti di dimensione nxq_i. Posto $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m]$, $\boldsymbol{u} = [\boldsymbol{u}_1', \ldots, \boldsymbol{u}_m']'$ e $q = \sum_{i=1}^{m} q_i$ il modello (2.2) può essere riscritto come il modello (2.1).

Le condizioni sufficienti per poter stimare i parametri sono:

- $\boldsymbol{u}$ ed $\boldsymbol{e}$ indipendenti
- $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma_0^2\boldsymbol{\Sigma}_e)$ e $\boldsymbol{u}_i \sim N_{q_i}(\boldsymbol{0}, \sigma_i^2\boldsymbol{\Sigma}_{u_i})$, con $\boldsymbol{\Sigma}_e$ e $\boldsymbol{\Sigma}_{u_i}$ note, $i = 1, \ldots, m$
- rango$(\boldsymbol{X}) = p$
- $n \geq p + m + 1$
- rango$(\boldsymbol{X} : \boldsymbol{Z}_i) > p$, $i = 1, \ldots, m$
- dato $\boldsymbol{V} = \sum_{i=0}^{m}\sigma_i^2\boldsymbol{G}_i$, con $\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_m$ linearmente indipendenti
- $\boldsymbol{Z}_i$ matrice di zero e uno tale che ci sia esattamente un 1 in ogni riga e al massimo un 1 in ogni colona, con $i = 1, \ldots, m$

Con queste condizioni, secondo il modello (2.2), $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$.

Per stimare i parametri incogniti $\boldsymbol{\beta}$, σ_0^2, σ_i^2, $i = 1, \ldots, m$ si può utilizzare il metodo della massima verosimiglianza, il metodo della massima verosimiglianza ristretta oppure il metodo dei momenti. Per un analisi più dettagliata sui modelli lineari ad effetti misti si veda Jiang (2007), Demidenko (2004), Goldstein (2003), McCullogh e Searle (2001).

I modelli lineari ad effetti misti si adattano naturalmente alla stima per piccole aree poiché permettono di stimare sia un effetto comune delle variabili ausiliarie sulla variabile di studio, sia un effetto per ogni area, utilizzando la stima degli effetti casuali. Si consideri in proposito il modello (2.2). Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita, $\boldsymbol{y} = (y_1, \ldots, y_N)'$ il vettore della variabile d'interesse (per tutte le unità della popolazione Ω). Si consideri un campione $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^m n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^m r_i = N - n$. Si consideri il modello (2.2) con $q = 1$, dove $\boldsymbol{Z}$ è una matrice di dimensione nxm con l'elemento della riga h colonna i uguale ad 1 se $h \in s_i$ e 0 altrimenti.

Si consideri uno stimatore generico $\hat{\theta}$ (ad esempio la media o il totale), combinazione lineare dei valori osservati e dei valori predetti con il modello (2.2)

$$\theta = \boldsymbol{a}_s'\boldsymbol{y}_s + \boldsymbol{a}_r'\boldsymbol{y}_r.$$

dove $\boldsymbol{a} = (\boldsymbol{a}_s', \boldsymbol{a}_r')'$ è un vettore di costanti note di dimensione Nx1. La stima $\hat{\theta}$ si ottiene stimando il modello (2.2) e ottenendo i valori predetti per $\boldsymbol{y}_r$, dunque $\hat{\theta} = \boldsymbol{a}_s'\boldsymbol{y}_s + \boldsymbol{a}_r'\hat{\boldsymbol{y}}_r$. Se l'ipotesi su y_i è quella formalizzata nel modello (2.2) la stima della media per piccola area è:

$$\hat{\bar{y}}_i = N_i^{-1}\left\{\sum_{j \in s_i} y_{ji} + \sum_{j \in r_i} \boldsymbol{x}_{ji}'\hat{\boldsymbol{\beta}} + \hat{u}_i\right\}$$

dove y_{ji} è la variabile di studio per l'unità j appartenente all'area i, N_i è il numero di unità della popolazione nell'area i, $\boldsymbol{x}_{ji}$ è il vettore delle variabili ausiliarie per l'unità j nell'area i, $\hat{\boldsymbol{\beta}}$ è la stima del vettore dei coefficienti di regressione (effetti fissi) e $\hat{u}_i$ è la stima dell'effetto casuale per l'area i, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$, $\hat{\boldsymbol{u}} = \boldsymbol{V}_u\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

La stima dell'errore quadratico medio (MSE, Mean Squared Error) si può ottenere sia tramite un'approssimazione al secondo ordine del polinomio di Taylor, sia con procedure di ricampionamento (bootstrap e jackknife). Il metodo più diretto è l'approssimazione al secondo ordine del polinomio di Taylor. L'errore quadratico medio (o MSE)[1] per $\hat{\theta}$ è

$$MSE(\hat{\theta}) \approx g_1(\boldsymbol{\sigma}) + g_2(\boldsymbol{\sigma}) + g_3(\boldsymbol{\sigma}) + g_4(\boldsymbol{\sigma})$$

dove

$$g_1(\boldsymbol{\sigma}) = \boldsymbol{a}_r'\boldsymbol{Z}_r\boldsymbol{T}_s\boldsymbol{Z}_r'\boldsymbol{a}_r,$$

$$g_2(\boldsymbol{\sigma}) = [\boldsymbol{a}_r'\boldsymbol{X}_r - \boldsymbol{a}_r'\boldsymbol{Z}_r\boldsymbol{T}_s\boldsymbol{Z}_s'\boldsymbol{V}_{es}'\boldsymbol{X}_s](\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{X}_s)^{-1}[\boldsymbol{X}_r'\boldsymbol{a}_r - \boldsymbol{X}_s'\boldsymbol{V}_{es}^{-1}\boldsymbol{Z}_s\boldsymbol{T}_s\boldsymbol{Z}_r'\boldsymbol{a}_r],$$

$$g_3(\boldsymbol{\sigma}) = tr\left\{(\nabla\boldsymbol{b}')\boldsymbol{V}_s(\nabla\boldsymbol{b}')'E[(\hat{\sigma} - \sigma)(\hat{\sigma} - \sigma)']\right\},$$

[1] Acronimo inglese per *Mean Squared Error*. Si useranno i due termini indifferentemente.

$g_4(\boldsymbol{\sigma}) = \boldsymbol{a}_r' \boldsymbol{V}_{er} \boldsymbol{a}_r,$

con $\boldsymbol{\sigma} = (\sigma_0^2, \sigma_i^2/\sigma_0^2, i = 1, \ldots, m)$, $\boldsymbol{b}' = (b_1, \ldots, b_n) = \boldsymbol{a}_r' \boldsymbol{Z}_r \boldsymbol{\Sigma}_u \boldsymbol{Z}_s' \boldsymbol{V}_s^{-1}$, $\boldsymbol{T}_s = \boldsymbol{\Sigma}_u - \boldsymbol{\Sigma}_u \boldsymbol{Z}_s' (\boldsymbol{\Sigma}_{es} + \boldsymbol{Z}_s \boldsymbol{\Sigma}_u \boldsymbol{Z}_s')^{-1} \boldsymbol{Z}_s \boldsymbol{\Sigma}_u$. Nel caso in cui $\hat{\sigma}$ sia uno stimatore approssimativamente corretto per $\boldsymbol{\sigma}$ si può ottenere una stima per l'errore quadratico medio di $\hat{\theta}$:

$$\hat{MSE}(\hat{\theta}) = g_1(\hat{\boldsymbol{\sigma}}) + g_2(\hat{\boldsymbol{\sigma}}) + 2g_3(\hat{\boldsymbol{\sigma}}) + g_4(\hat{\boldsymbol{\sigma}}).$$

Per una trattazione completa della stima per piccole aree con modelli lineari ad effetti misti si veda Rao (2003).

2.3 Il modello di regressione M-quantile applicato alla stima per piccole aree

Il modello di regressione M-quantile nasce, anche, come alternativa al modello di regressione quantilica. Lo scopo è, come per la regressione quantilica, quello di modellare le "code" della distribuzione di $Y|X$ (la variabile di studio date le covariate). L'articolo di Breckling e Chambers (1988) sul modello di regressione M-quantile, oltre che riferirsi alla regressione quantilica (Koenker e Bassett (1978)), si basa sull'articolo di Newey e Powell (1987) a proposito degli *Asymmetric Least Squares* e sul cosidetto *M-estimator* proposto da Huber (1981).

Prima di presentare il modello di regressione M-quantile si ritiene utile una nota metodologica sul modello di regressione quantilica, sugli asymmetric least sqaures e sugli M-estimator.

Nell'ultima sezione di questo paragrafo si mostrerà come applicare il modello M-quantile alla stima per piccole aree.

2.3.1 La regressione quantilica

Introduzione alla regressione quantilica

Citando Mosteller e Tukey (1977):

> «la regressione (lineare) esprime in modo completo il comportamento in media della distribuzione di una certa variabile, Y, condizionatamente ad un insieme di altre variabili, X. E' possibile andare oltre e calcolare diverse curve di regressione corrispondenti a diversi percentili della distribuzione condizionata per avere un quadro più completo sui dati. Normalmente ciò non viene fatto, per questo la regressione offre spesso un quadro incompleto dei dati. Proprio come la media offre un quadro incompleto di una singola distribuzione anche la regressione offre un corrispondente quadro incompleto sulla distribuzione condizionata di Y dato X».

L'obiettivo della regressione quantilica è proprio quello di completare "il quadro", ovvero studiare il comportamento di Y condizionato a X non solo in media (regressione lineare) ma anche rispetto a determinati quantili.

La logica con cui si giunge all'equazione di stima della regressione quantilica è semplice (Koenker e Bassett (1978)): è noto che la media aritmetica μ è la soluzione del problema

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Se si esprime la media condizionata di $Y|X$ come $\mu(X) = \boldsymbol{X}'\boldsymbol{\beta}$ allora si può stimare $\boldsymbol{\beta}$ risolvendo il sistema

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Questo problema di minimo dà luogo alle note equazioni dei minimi quadrati. Procedendo analogamente, per i quantili di una distribuzione si ha che il τ-esimo quantile, $q(\tau)$, è la soluzione del problema

$$\min_{\alpha\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - q),$$

dove $\rho_\tau(t) = t\,(\tau - I(t < 0))$. Esprimendo un certo quantile della distribuzione di $Y|X$ come $Q_y(\tau|x) = \boldsymbol{X}'\boldsymbol{\beta}(\tau)$, dunque linearmente dipendente dalla X, si ottiene l'equazione di stima per $\boldsymbol{\beta}(\tau)$ risolvendo il problema

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}).$$

Il processo di stima nella regressione quantilica

Sia Y la variabile casuale di studio e sia $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ una sua realizzazione. Sia $\boldsymbol{X}$ una matrice di dimensioni nxp, dove $\boldsymbol{x}_{ik}'$, $i = (1, \ldots, n)$, $k = (1, \ldots, p)$, è il vettore della k-esima variabile ausiliaria che corrisponde alla k-esima colonna della matrice $\boldsymbol{X}$.

Come accennato nell'introduzione, l'obiettivo è quello di modellare i quantili della distribuzione condizionata di Y dato X (Koenker e Bassett (1978)). Tale modellazione è nota con il nome di *regressione quantilica* (o in inglese *quantile regression*). Il modello di regressione quantilico può essere lineare o non lineare. Si consideri il modello lineare (quello presentato precedentemente) per il τ-esimo quantile della distribuzione condizionata di Y dato X:

$$Q_y(\tau|\boldsymbol{x}) = \boldsymbol{X}'\boldsymbol{\beta}(\tau).$$

Ricordando il processo logico discusso precedentemente, il problema della stima del vettore dei coefficienti di regressione, $\boldsymbol{\beta}(\tau)$, si riconduce alla minimizzazione della seguente funzione:

$$R(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.3}$$

dove $\rho_\tau(t) = t\,(\tau - I(t < 0))$, con $I(\cdot)$ funzione indicatrice e $\boldsymbol{\beta} \in \mathbb{R}^p$.

La (2.3) è lineare e continua a tratti ed è derivabile ad eccezione dei punti in cui il residuo, $y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, è pari a zero. In questi punti particolari $R(\beta)$ ha le derivate direzionali che "puntano" in infinite direzioni rispetto alla direzione che si vuole valutare. La derivata direzionale di $R(\beta)$ in una certa direzione $\boldsymbol{w}$ è:

$$\begin{aligned} \nabla R(\beta, w) &= \frac{d}{dk} R(\beta + t\boldsymbol{w})|_{k=0} \\ &= -\sum \psi_\tau^* \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}; -\boldsymbol{x}_i'\boldsymbol{w}\right) \boldsymbol{x}_i'\boldsymbol{w} \end{aligned}$$

dove

$$\psi_\tau^*(t,v) = \begin{cases} \tau - I(t<0) & (t \neq 0) \\ \tau - I(v<0) & (v=0) \end{cases}$$

Se in un dato punto $\hat{\boldsymbol{\beta}}$ le derivate direzionali sono tutte non negative, cioè $\nabla R(\hat{\boldsymbol{\beta}}, \boldsymbol{w}) \geq 0 \forall \boldsymbol{w} \in \mathbb{R}^p$ con $||\boldsymbol{w}|| = 1$, allora $\hat{\boldsymbol{\beta}}$ minimizza la funzione obiettivo (2.3). La soluzione trovata, $\hat{\boldsymbol{\beta}}$, corrisponde ai punti nello spazio parametrico dove p osservazioni sono interpolate quando si stimano p parametri. Questo non significa che si utilizzano solo "p" osservazioni per ottenere le stime ma significa che vengono identificate p osservazioni; infatti come la mediana identifica l'osservazione che divide in due parti uguali una distribuzione, così la regressione quantilica (preso ad esempio $\tau = 0.5$) identifica p osservazioni che definiscono un iperpiano che rappresenta al meglio il quantile (mediana) nella distribuzione condizionata alle X.

La minimizzazione della funzione (2.3) non sempre è poossibile. Esistono delle condizioni che garantiscono l'esistenza di una soluzione. Si consideri $h = (1, \ldots, p) \in \mathcal{H}$, dove $\mathcal{H}$ è un sottoinsieme dei primi n numeri naturali, $\mathcal{N} = \{1, 2, \ldots, n\}$. Inoltre sia $\boldsymbol{X}(h)$ la sottomatrice di $\boldsymbol{X}$ con righe $\{\boldsymbol{x}_i : i \in h\}$ e $\boldsymbol{y}(h)$ un vettore di dimensione p con elementi $\{y_i : i \in h\}$. Utilizzando la notazione introdotta, ogni soluzione del problema di minimizzazione di $R(\boldsymbol{\beta})$ che passa per i punti $\{(x_i, y_i) : i \in h\}$ è definita *soluzione base*:

$$\boldsymbol{b}(h) = \boldsymbol{X}(h)^{-1}\boldsymbol{y}(h),$$

considerando che $\boldsymbol{X}(h)$ sia non singolare. Sfortunatamente esiste un numero eccessivamente grande di soluzioni di base per confrontarle tutte e trovare la soluzione ottima (Koenker, 2005). Grazie alle tecniche di programmazione lineare (come l'algoritmo del simplesso) è possibile ottenere una soluzione ottima per la (2.3).

Si consideri la seguente definizione di Rousseeuw e Leroy (1987):

Definizione 2.3.1 *le osservazioni $(y_i, \boldsymbol{x}_i)$, con $i = 1, \ldots, n$, si definiscono* general position *se p di esse producono un unico adattamento perfetto del modello di regressione quantilica per un certo quantile τ, ovvero per ogni $h \in \mathcal{H}$ si verifica $y_i - \boldsymbol{x}_i'\boldsymbol{b}(h) \neq 0 \forall i \notin h$.*

Utilizzando la definizione 2.3.1 Koenker (2005) presenta un teorema per le condizioni necessarie per l'esistenza di una soluzione ottima per la (2.3):

Teorema 2.3.1 *se $(y_i, \boldsymbol{x}_i)$ sono general position, allora esiste una soluzione per la* (2.3) *della forma $\boldsymbol{b}(h) = \boldsymbol{X}(h)^{-1}\boldsymbol{y}(h)$ se e solo se per qualche $h \in \mathcal{H}$ si verifica*

$$-\tau 1_p \leq \xi(h) \leq (1-\tau)1_p, \tag{2.4}$$

dove $\xi'(h) = \sum_{i \in h} \psi_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{b}(h))\boldsymbol{x}_i'\boldsymbol{X}(h)^{-1}$, con $\psi_\tau(t) = \tau - I(t < 0)$. Inoltre $\boldsymbol{b}(h)$ è una soluzione unica se e solo se la diseguaglianza nella (2.4) *vale solo strettamente.*

Nonostante l'idea di studiare il comportamento della distribuzione di $Y|X$ per certi quantili sia molto attraente ci si scontra con la difficoltà oggettiva di trovare una stima ottima per $\boldsymbol{\beta}$. E' opportuno sottolineare che la definizione 2.3.1 si riferisce a casi frequenti e dunque il problema della stima non è determinante per la sfruttabilità di questo modello (Koenker (2005)).

Regressione Quantilica: interpretazione del modello

Nell'analisi tradizionale di regressione lineare, dove $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta}$, i coefficienti β_j, $j = (1, \ldots, p)$, sono le derivate parziali di $E[Y|\boldsymbol{X} = \boldsymbol{x}]$:

$$\frac{\partial E[Y|\boldsymbol{X} = \boldsymbol{x}]}{\partial x_j} = \beta_j,$$

in pratica β_j indica il variare di Y in media rispetto ad una variazione unitaria in x_j, fermo restando tutte le altre variabili ausiliarie. Il problema d'interpretazione di β_j è meno diretta nel caso di trasformazioni della Y: $E[h(Y)|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta}$. In un caso del genere si presume che l'interesse sia comunque rivolto al comportamento della Y condizionata alle X ($E[Y|\boldsymbol{X} = \boldsymbol{x}]$). Intuitivamente si potrebbe ipotizzare che

$$E[h(Y)|\boldsymbol{X} = \boldsymbol{x}] = h(E[Y|\boldsymbol{X} = \boldsymbol{x}]). \tag{2.5}$$

Purtroppo questa relazione non vale in generale, ha valore solo in casi particolari. Si prenda come esempio la diffusa trasformazione $h(Y) = \log(Y)$. Spesso si ipotizza la relazione (2.5) per l'interpretazione dei risultati, ma tale pratica è errata. Il problema nasce dal fatto che $E[h(Y)]$ non sempre è uguale a $h(E[Y])$; questo rende l'interpretazione del modello di regressione lineare complicata, in certe circostanze.

Nel caso della regressione quantilica il problema appena affrontato è completamente ridimensionato. Si consideri la regressione quantilica per il τ-esimo quantile della variabile di studio. L'interpretazione dei coefficienti è tale e quale a quella della della regressione lineare: dato $Q_y(\tau|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}(\tau)$ si interpretano i coefficienti in base alle derivate parziali

$$\frac{\partial Q_y(\tau|\boldsymbol{X} = \boldsymbol{x})}{\partial x_j} = \beta_j(\tau).$$

A differenza della regressione lineare, se si deve modellare $h(Y)$, con $h(\cdot)$ monotona, con la regressione quantilica risulta

$$Q_{h(y)}(\tau|\boldsymbol{X} = \boldsymbol{x}) = h(Q_y(\tau|\boldsymbol{X} = \boldsymbol{x})).$$

L'interpretazione dei coefficienti per questo tipo di trasformazioni (monotone) rimane invariata, infatti

$$\frac{\partial Q_y(\tau|\boldsymbol{X} = \boldsymbol{x})}{\partial x_j} = \frac{\partial h^{-1}(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_j}.$$

Nel modello di regressione quantilica si ottiene in modo corretto l'effetto sulla variabile Y di una covariata, partendo da una qualsiasi trasformazione monotona. Per esempio, sia $h(Y) = \log(Y)$, il modello di regressione quantilica è $Q_{\log(y)}(\tau|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}(\tau)$. $\beta_j(\tau)$ è l'effetto sul τ-esimo quantile di $\log(Y)$ dell'incremento di una unità della variabile ausiliaria x_j. Per ottenre l'effetto di x_j sul τ-esimo quantile di Y basta derivare la funzione inversa del logaritmo:

$$\frac{\partial Q_y(\tau|\boldsymbol{X} = \boldsymbol{x})}{\partial x_j} = \frac{h^{-1}\boldsymbol{x}'\boldsymbol{\beta}(\tau)}{\partial x_j} = \beta_j(\tau)\exp(\boldsymbol{x}'\boldsymbol{\beta}(\tau)).$$

L'interpretazione dei risultati va comunque affrontata con la massima cautela. Si consideri l'effetto $\beta_j(\tau)$, esso indica che per $\Delta x_j = 1$ il τ-esimo quantile della variabile Y aumenta di $\beta_j(\tau)$ volte; se per un'osservazione appartenente al τ-esimo quantile risultasse $x_j = k$, non sarebbe detto

che l'osservazione $x_j = k + 1$ cada sempre nel τ-esimo quantile, per il quale vale l'effetto $\beta_j(\tau)$. Un'osservazione con valore di $x_j = k + 1$ potrebbe appartenere ad una altro quantile, per esempio $\tau + \epsilon$, per il quale vale l'effettto $\beta_j(\tau + \epsilon) \neq \beta_j(\tau)$. L'interpretazione dei risultati del modello di regressione quantilica, anche se è facile e versatile da un punto di vista matematico, va affrontata con particolare attenzione e cura a causa del problema appena esposto.

Un altro problema rilevante che merita attenzione è noto come *quantile crossing*. Per quantile crossing si intende la possibilità che due o più rette (o iperpiani, nel caso di più regressori) si incrocino. Da un punto di vista interpretativo questo evento è assurdo: il valore di un quantile di un certo ordine risulterebbe superiore al valore di un quantile di ordine superiore. Per esempio può accadere che dato $\tau_1 < \tau_2$ si verifichi $\hat{Q}_y(\tau_1|\boldsymbol{X} = \boldsymbol{x}) > \hat{Q}_y(\tau_2|\boldsymbol{X} = \boldsymbol{x})$, dove $\hat{Q}_y(\tau_i|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}\hat{\boldsymbol{\beta}}(\tau_i)$. In questo caso si viola il principio della non decrescenza della funzione di ripartizione e della sua inversa, che definisce i quantili. Il quantile crossing si può verificare all'interno dello spazio campionario o all'esterno. Se si verifica all'esterno si può considerare valido il modello nel solo spazio campionario. Altrimenti si dovrebbe riconsiderare il modello utilizzato poiché il quantile crossing potrebbe essere indice di eteroschedasticità, multicollinearità o errata specificazione del modello. Tuttavia il problema del quantile crossing è confinato nelle "estremità" dello spazio campionario. Si consideri a riguardo il seguente teorema (Koenker (2005)):

Teorema 2.3.2 *sia* $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. $\hat{Q}_y(\tau|\bar{x})$ *è non decrescente in* τ, *con* $\tau \in [0, 1]$.

Nel centroide dello spazio campionario ($\bar{x}$) la stima del quantile della distribuzione condizionata di $Y|X$ è monotona in τ. Koenker (2005) sottolinea che il teorema 2.3.2 non risolve il problema del quantile crossing, tuttavia se il modello lineare è correttamente specificato il problema del quantile crossing dovrebbe presentarsi raramente. Il problema del quantile crossing è stato affrontato anche da He (1997), che ha presentato una versione "vincolata" della regressione quantilica che garantisce l'assenza del fenomeno[2].

Conclusioni sul modello di regressione quantilica

La regressione quantilica offre la possibilità di conoscere tutta la distribuzione di $Y|X$. Per la teoria asintotica della stima, il test di ipotesi, gli intervalli di confidenza, la stima di varianza e covarianza e ulteriori argomenti legati alla regressione quantilica si consulti Koenker (2005); l'autore non tralascia neanche i modelli quantilici non lineari e non parametrici. Nonostante il modello offra notevoli vantaggi, soprattutto rispetto a trasformazioni lineari della variabile di studio, presenta almeno due punti deboli di notevole rilevanza: **i.** nella procedura di stima non è garantita l'ottimalità dei parametri, **ii.** il quantile crossing. Nell'ambito della stima per piccole aree il quantile crossing, se si verifica al di fuori dello spazio campionario, non è un problema determinante poiché i modelli hanno una funzione predittiva e non esplicativa. Il primo punto, al contrario, è un problema determinante, e non solo nell'ambito della stima per piccole aree. Tuttavia, la spinta verso la ricerca di altri modelli deriva, probabilmente, dal fatto che il modello di regressione quantilica non si presta a "confronti" diretti con il modello di regressione lineare. Questo non significa che non si può confrontare l'effetto medio con l'effetto mediano o di altri quantili, significa però che tale confronto non è fatto sullo stesso "piano"; la media e la mediana sono due concetti differenti. E' stato ritenuto opportuno creare un metodo in grado di modellare le code della distribuzione restando sullo stesso "piano" della media, cioè del $E[Y|X]$. Tale metodo è noto come *asymmetric least squares*, (Newey e Powell, 1987) come accenato all'inizio di questo paragrafo.

[2] *Restricted Regression Quantile*.

2.3.2 Asymmetric Least Squares

Introduzione

Il metodo degli Asymmetric Least Squares (ALS) è stato proposto da Newey e Powell (1987), che si sono basati su un lavoro precedente di Aigner *e altri* (1976). Newey e Powell (1987) utilizzano gli ALS per identificare eteroschedasticità e asimmetria in un modello di regressione lineare. L'idea alla base degli ALS è la stessa della regressione quantilica, ma è implementata utilizzando un approccio che rende la stima dei parametri più facile da calcolare. Il risultato è quello di poter modellare interamente la distribuzione condizionata di $Y|X$ (la variabile di interesse dato le covariate), ma in modo che la modellazione del "centro" della distribuzione corrisponda alla modellazione del $E[Y|X]$. Ad esempio se si utilizza un modello di regressione lineare per $E[Y|X]$ allora il caso particolare del "centro" della distribuzione di $Y|X$ modellata con gli ALS corrisponderà al modello di regressione lineare stesso.

In questo paragrafo si considera solo il caso in cui gli errori del modello di regressione lineare siano indipendenti e identicamente distribuiti (Aigner *e altri* (1976)), ciò non toglie che gli stimatori ALS siano validi anche per casistiche più generali (errori correlati ed eteroschedasticità, per fare un esempio).

Asymmetric Least Squares: definizione e stima

Si consideri che le coppie di valori osservabili $\{(y_i, \boldsymbol{x}_i'), i = 1, \ldots, n\}$ siano generate dal seguente modello:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i \tag{2.6}$$

dove $\boldsymbol{x}_i$ è il vettore di dimensione p delle covariate per l'i-esima osservazione, $\boldsymbol{\beta}_0$ è il vettore di dimensione p dei coefficienti di regressione, y_i è il valore della variabile di interesse per l'unità i-esima e ε_i è l'errore associato all'osservazione i-esima.

Analogamente a quanto avviene nella regressione quantilica e a quanto avviene nella regressione lineare, si determina una funzione dell'errore del modello di regressione da minimizzare. Nella regressione lineare classica si minimizza la funzione $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$ mentre per la regressione quantilica (nella sua forma lineare) la funzione $\sum_{i=1}^{n}|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}||\tau - I(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} < 0)|$, per un quanitle di ordine τ. Generalizzando il problema, si identifica una funzione generica da minimizzare:

$$\sum_{i=1}^{n}\rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}),$$

dove ρ_τ è detta funzione di perdita. Nel caso della regressione lineare classica la funzione di perdita è $\rho_\tau(t) = \rho(t) = t^2$ mentre per la regressione quantilica è $\rho_\tau(t) = |\tau - I(t < 0)||t|$. Nel caso degli ALS la funzione di perdita è una fusione delle funzioni di perdita della regressione quantilica e della regressione classica.

Formalizzando il problema, il sistema di equazioni di stima da minimizzare per gli ALS è:

$$\sum_{i=1}^{n}\rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = \min, \tag{2.7}$$

con funzione di perdita

$$\rho_\tau(t) = |\tau - I(t < 0)|t^2, \tag{2.8}$$

dove $\tau \in (0,1)$ è il fattore che determina la "posizione" nella distribuzione di interesse[3]. La (2.7) si minimizza rispetto a $\boldsymbol{\beta}$ dato un certo valore di τ.

$\boldsymbol{\beta}(\tau)$ è il vettore di parametri che determinano l'effetto delle covariate sulla variabile Y in una data posizione, determinata da τ, della distribuzione di Y dato X. Per ottenere una stima di $\boldsymbol{\beta}(\tau)$ si consideri il parametro $\mu(\tau)$ ottenuto minimizzando la funzione $E[\rho_\tau(Y-m)-\rho_\tau(Y)]$ rispetto a m, dove il valore atteso è riferito alla distribuzione della variabile casuale Y, per la quale si assume che la media sia finita. Newey e Powell (1987) mostrano che il parametro $\mu(\tau)$ è la soluzione dell'equazione

$$\mu(\tau) - E[Y] = \left[(2\tau-1)(1-\tau)^{-1}\right] \int_{\mu(\tau)}^{\infty} (y-\mu(\tau))\, dF(y), \tag{2.9}$$

dove $F(y)$ è la funzione di ripartizione di Y. Un'equazione alternativa per $\mu(\tau)$, suggerita da A. Goldberg[4] è

$$\tau(1-\tau)^{-1} = \left[\int_{-\infty}^{\mu(\tau)} (\mu(\tau)-y) dF(y)\right] \left[\int_{\mu(\tau)}^{\infty} (\mu(\tau)-y) dF(y)\right]^{-1}. \tag{2.10}$$

Confrontando questa equazione con la relazione analoga per i quantili, ovvero

$$\tau(1-\tau)^{-1} = F(q(\tau))(1-F(q(\tau)))^{-1} = \left[\int_{-\infty}^{q(\tau)} dF(y)\right] \left[\int_{q(\tau)}^{\infty} dF(y)\right]^{-1}, \tag{2.11}$$

si noti come la (2.10) esprima lo stesso concetto della (2.11) utilizzando però il valore atteso applicato ad una data posizione della distribuzione. Quindi $\mu(\tau)$ è determinato dal valore atteso della variabile casuale Y condizionata dal fatto che Y sia nella coda (cioè in una posizione diversa dal centro ed espressa da $\tau \in [0,1]$, $\tau \neq 0.5$) della distribuzione. Per questo motivo si fa riferimento a $\mu(\tau)$ con il nome di *expectile*. Utilizzando l'expectile $\mu(\tau)$ si può caratterizzare completamente la funzione di distribuzione, esattamente come avviene nel caso dei quantili, $q(\tau) = F^{-1}(\tau)$.

Per trovare il valore di $\mu(\tau)$ si deve risolvere la (2.9). La soluzione della (2.9) e garantita dal seguente teorema di Newey e Powell (1987):

Teorema 2.3.3 *si ipotizzi che $E[Y] = m$ esista, con m finito. Allora per ogni $0 < \tau < 1$ esiste un'unica soluzione $\mu(\tau)$ all'equazione* (2.9) *ed ha le seguenti proprietà:*

i. *$\mu(\tau) : (0,1) \to \mathbb{R}$, $\mu(\tau)$ è una funzione strettamente monotona crescente.*

ii. *Il range di $\mu(\tau)$ è uguale al supporto della variabile casuale Y e $\mu(\tau)$ mappa $(0,1)$ nel supporto di Y.*

iii. *Per $\tilde{Y} = sY + t$, dove $s > 0$, il τ^o expectile $\tilde{\mu}(\tau)$ di $\tilde{Y}$ soddisfa l'uguaglianza $\tilde{\mu}(\tau) = s\mu(\tau) + t$.*

iv. *Se $F(y)$ è continua e differenziabile, allora $\mu(\tau)$ è continua e differenziabile, e per $y \neq m$ e τ_y tale che $y = \mu(\tau_y)$, vale*

$$F(y) = -(y - m + \tau_y \mu'(\tau_y)(1-2\tau_y))(\mu'(\tau_y)(1-2\tau_y)^2)^{-1},$$

che vale in limite anche per $y = m$ e $\tau_y = 1/2$ ($\mu'(\tau_y)$ è la derivata prima di $\mu(\tau_y)$).

[3] Si utilizza il termine "posizione" poiché τ in questo caso non è un quantile, anche se è un suo analogo.

[4] Che si trova in nota nell'articolo di Newey e Powell (1987)

La proprietà **iii.** indica che, come per il τ-esimo quantile, il τ-esimo expectile è equivariante rispetto a traslazioni e cambiamenti di scala. Le proprietà più importanti, però, sono la **ii.** e la **iv**: insieme implicano che la funzione $\mu(\tau)$ caratterizza interamente la distribuzione di Y (come precedentemente accennato).

Si è visto come l'expectile abbia caratteristiche simili ai quantili. Al fine di verificare la differenza tra l'expectile e il quantile si consideri la figura 2.1 in cui sono riportate le funzioni expectile e quantile per una distribuzione normale standard.

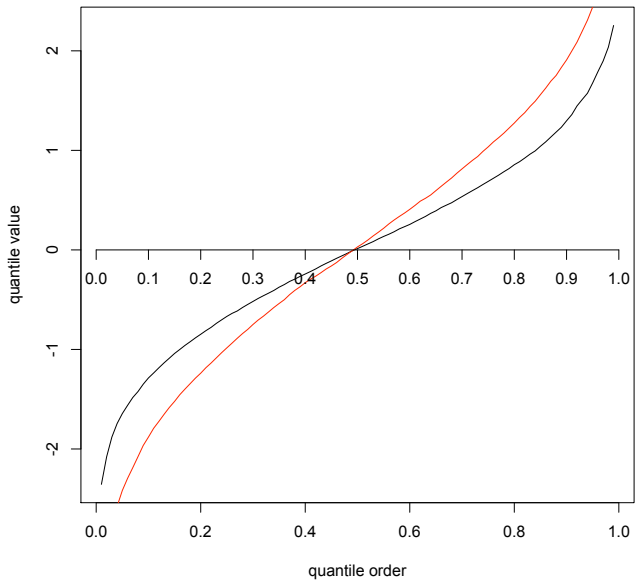


Figura 2.1: Funzione quantile $q(\tau)$ in rosso e expectile $\mu(\tau)$ in nero per una distribuzione normale standard.

Dalla figura 2.1 si vede che la funzione expectile ha una pendenza minore rispetto alla funzione quantile nell'intorno di $\tau = 0.5$, mentre ha una pendenza maggiore nell'intorno di $\tau = 0$ e $\tau = 1$. Anche considerando un'altra distribuzione nota, come la uniforme continua tra 0 e 1, la funzione expectile ha un comportamento simile a quello appena descritto, rispetto alla funzione quantile.

Newey e Powell (1987), una volta definito l'expectile, propongono un suo utilizzo nella regressione lineare (e non lineare). Si consideri il τ-esimo expectile di y_i condizionato da $\boldsymbol{x}_i$, $\mu(\tau, \boldsymbol{x}_i)$, che si ottiene minimizzando $E[\rho_\tau(y_i - m) - \rho_\tau(y_i)|\boldsymbol{x}_i]$ rispetto a m, dove il valore atteso è riferito alla distribuzione condizionata $Y|X$. Sia $\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)$ l'approssimazione lineare, in $\boldsymbol{x}_i$, dell'expectile $\mu(\tau, \boldsymbol{x}_i)$. Estendendo il teorema 2.3.3 per la regressione, Newey e Powell (1987) mostrano che l'expectile condizionato $\mu(\tau, \boldsymbol{x}_i)$ caratterizza in modo completo la distribuzione condizionata di y_i dato $\boldsymbol{x}_i$. Si è detto che $\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)$ è un'approssimazione lineare dell'expectile condizionato, quindi utilizzando tale approssimazione si può caratterizzare la distribuzione condizionata $Y|X$ (anche se in modo approssimato). Il vettore dei coefficienti di regressione $\boldsymbol{\beta}(\tau)$, per l'approssimazione dell'expectile $\mu(\tau, \boldsymbol{x}_i)$, si ottiene minimizzando la funzione $E[\rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) - \rho_\tau(y_i)]$ che dipende dalla distribuzione condizionata di y_i dato $\boldsymbol{x}_i$. Newey e Powell (1987) dimostrano che $\boldsymbol{\beta}(\tau)$ è la soluzione dell'equazione

$$\boldsymbol{\beta}(\tau) = \left(E[|\tau - I(y_i < \boldsymbol{x}_i'\boldsymbol{\beta}(\tau))|\boldsymbol{x}_i y_i]\right)\left(E[|\tau - I(y_i < \boldsymbol{x}_i'\boldsymbol{\beta}(\tau))|\boldsymbol{x}_i\boldsymbol{x}_i']\right)^{-1}.$$

La relazione tra $\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)$ e $\mu(\tau, \boldsymbol{x}_i)$ dipende dal modello relazionale utilizzato per la Y. Nel caso del modello (2.6) e ipotizzando che ε_i sia indipendente da $\boldsymbol{x}_i$, si ha che $\mu(\tau, \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + \mu(\tau)$ (che si ottiene dalla proprietà **iii.** del teorema 2.3.3), dove $\mu(\tau)$ è il τ-esimo expectile di ε_i. Poiché si è ipotizzato che l'expectile condizionato, $\mu(\tau, \boldsymbol{x}_i)$, è approssimato linearmente da $\boldsymbol{x}_i\boldsymbol{\beta}(\tau)$, segue che $\boldsymbol{\beta}(\tau)$, il vettore dei coefficienti di regressione, è:

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0 + \mu(\tau)e_1, e_1 = (1, 0, \ldots, 0)'. \tag{2.12}$$

Cambiando il valore di τ si ha un cambiamento solo nell'intercetta, che è il primo elemento del vettore $\boldsymbol{\beta}(\tau)$ (e del vettore $\boldsymbol{\beta}_0$).

In generale la relazione tra l'expectile condizionato, $\mu(\tau, \boldsymbol{x}_i)$, e $\boldsymbol{x}_i\boldsymbol{\beta}(\tau)$ può assumere forme più complesse, per esempio in caso di eteroschedasticità la relazione (2.12) coinvolgerebbe altri regressori oltre l'intercetta. In questi casi il modo di procedere alla determinazione di $\boldsymbol{\beta}(\tau)$ è del tutto analogo a quanto fatto precedentemente, fatta salva qualche complicazione algebrica.

Uno dei problemi più rilevanti della regressione quantilica è l'algoritmo di stima dei parametri. Per gli ALS questo problema non esiste. La funzione di perdita degli ALS, $\rho_\tau(t)$ (2.8), è continua e differenziabile in t, quindi si può ottenere lo stimatore $\hat{\boldsymbol{\beta}}(\tau)$ utilizzando il metodo dei minimi quadrati pesati iterati:

$$\hat{\boldsymbol{\beta}}(\tau) = \left[\sum_{i=1}^{n} |\tau - I(y_i < \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau))|\boldsymbol{x}_i y_i\right]\left[\sum_{i=1}^{n} |\tau - I(y_i < \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau))|\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1}.$$

Un piccolo svantaggio degli ALS rispetto alla regressione quantilica si ha nella difficile interpretazione dell'expectile. Tuttavia, come si vedrà in seguito, questo non è un particolare molto rilevante dal punto di vista dell'applicazione di questo strumento alla stima per piccole aree. Inoltre dal punto di vista dell'interpretazione Newey e Powell (1987) suggeriscono di calcolare la proporzione di osservazioni per cui $y_i < \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau)$ in modo da avere un'idea della posizione dello stimatore $\hat{\boldsymbol{\beta}}(\tau)$ nella distribuzione condizionata di Y dato X. Nel caso in cui ε_i sia indipendente da $\boldsymbol{x}_i$ tale proporzione è una stima consistente di $G(\mu(\tau))$, dove G è la funzione di ripartizione di ε_i e $\mu(\tau)$ è il τ-esimo expectile di ε_i.

Per intervalli di confidenza asintotici, stima della varianza-covarianza asintotica dello stimatore $\hat{\boldsymbol{\beta}(\tau)}$ e test di ipotesi sui parametri, su asimmetria ed eteroschedasticità si consulti Newey e Powell (1987).

2.3.3 M-Estimator

Introduzione

L'*M-estimator* è uno stimatore robusto di massima verosimiglianza (*M-estimator* è l'abbreviazione di *Maximum likelihood type estimator*).

La statistica inferenziale "classica" si basa, oltre che sui dati osservati, su assunzioni a priori circa la distribuzione di certi parametri inerenti il problema studiato. Si pensi, ad esempio, alle assunzioni che si fanno per trovare un intervallo di confidenza per la media (variabili indipendenti e identicamente distribuite). E' possibile che in un problema di inferenza venga affrontato in base ad ipotesi non vere, con la conseguenza, a volte, di ottenere risultati "distanti" dalla realtà dei fatti. Per ovviare a queste situazioni si è sviluppato un filone di ricerca noto come *statistica robusta*.

In statistica robustezza significa insensibilità a piccole deviazioni dalle assunzioni (Huber (1981)). Le deviazioni dalle assunzioni sono dovute principalmente a due cause: presenza di outliers e distribuzione reale diversa da quella ipotizzata nelle assunzioni. Nel primo caso si parla di robustezza verso gli outliers, nel secondo di robustezza rispetto alla distribuzione. Un esempio di stimatore robusto in entrambi i sensi è la mediana.

La robustezza si formalizza attraverso due aspetti, robustezza qualitativa e robustezza quantitativa.

La robustezza qualitativa è così definita:

Definizione 2.3.2 *Siano x_i osservazioni indipendenti e identicamente distribuite, con funzione di ripartizione F. Sia (T_n) una sequenza di stime, o di statistiche test, funzione delle x_i, $T_n = T_n(x_1, \ldots, x_n)$. Allora la sequenza (T_n) è detta* robusta in $F = F_0$ *se presa una funzione distanza d_* nello spazio di probabilità $\mathcal{M}$ e assunto che per ogni $\varepsilon > 0$, esiste un $\delta > 0$ e un $n_0 > 0$ tali che, per ogni F e ogni $n \geq n_0$,*

$$d_*(F_0, F) \leq \delta \Rightarrow d_*(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)) \leq \varepsilon$$

dove $\mathcal{L}_F$ è una funzione che mappa (T_n) in F.

L'aspetto quantitativo, invece, misura quanto un piccolo cambiamento nella distribuzione sottostante ai dati, F, cambi la distribuzione, $\mathcal{L}_F(T_n)$, di una stima o di una statistica test, $T_n = T_n(x_1, \ldots, x_n)$. In pratica si cerca di misurare l'impatto sugli stimatori di assunzioni non vere.

Stimatori del tipo di massima verosimiglianza (M-Estimator)

Si consideri la funzione di $\log$ massima verosimiglianza per la funzione di densità di una variabile casuale X dipendente dal parametro θ, $f(x_i, \theta)$:

$$\sum_{i=1}^{n} \log f(x_i, \theta) = \max,$$

che equivale a

$$\sum_{i=1}^{n} - \log f(x_i, \theta) = \min .$$

Utilizzando una funzione di perdita $\rho(x_i, \theta) = -\log f(x_i, \theta)$ la funzione di massima verosimiglianza diventa:

$$\sum_{i=1}^{n} \rho(x_i, \theta) = \min . \tag{2.13}$$

Con la (2.13) si può esprimere una gamma di stimatori di cui lo stimatore di massima verosimiglianza è un caso particolare. Si consideri la seguente definizione di M-estimator (Huber, 1981):

Definizione 2.3.3 *Ogni stima T_n definita da un problema di minimo del tipo*

$$\sum \rho(x_i; T_n) = \min$$

o da un'equazione

$$\sum \psi(x_i; T_n) = 0,$$

dove ρ è una qualunque funzione tale che $\psi(x;u) = \partial\rho(x;u)/\partial u$, è un M-estimator.

Infatti posto $\rho(x_i, T_n) = -\log f(x_i, T_n)$, dalla definizione 2.3.3 si ottiene l'equazione di stima di massima verosimiglianza tradizionale per il parametro T_n.

E' noto che lo stimatore media aritmetica si ottiene minimizzando $\sum(x_i - T_n)^2$ rispetto a T_n, dunque una gamma di stimatori robusti sono del tipo:

$$\sum \rho(x_i - T_n) = \min,$$

o

$$\sum \psi(x_i - T_n) = 0. \tag{2.14}$$

T_n in questo caso è una stima di posizione. La (2.14) può essere riscritta utilizzando dei pesi:

$$\sum w_i(x_i - T_n) = 0,$$

con

$$w_i = (\psi(x_i - T_n))(x_i - T_n)^{-1}$$

da cui si può riscrivere T_n come media ponderata, $T_n = (\sum w_i x_i)(\sum w_i)^{-1}$.

Per le proprietà asintotiche degli stimatori M-estimator e gli aspetti di robustezza qualitativa e quantitativa per questa gamma di stimatori si consulti Huber (1981).

Nell'utilizzo degli M-estimator gioca un ruolo fondamentale la scelta della funzione ρ (o ψ) che è legata alla funzione di influenza. La funzione di influenza, $IC(x, F, T)$, di un M-estimator è:

$$IC(x, F, T) = (\psi(x; T(F))) \left(- \int (\partial\psi(x; T(F))/\partial\theta) F(dx) \right)^{-1},$$

che in un problema per la stima di parametri posizionali assume la seguente forma:

$$IC(x, F, T) = (\psi(x - T(F))) \left(\int (\partial\psi(x - T(F))/\partial\theta) F(dx) \right)^{-1},$$

dove θ è il parametro (o vettore di parametri) che caratterizza la funzione di ripartizione F e $T(F)$ indica uno stimatore consistente ($T_n \to T(F)$). La scelta della funzione di influenza viene fatta in base all'efficienza asintotica delle stime. Sia $(F_\theta)_{\theta\in\Theta}$ una famiglia di distribuzioni dipendenti dal parametro $\theta \in \Theta$, doveΘ è lo spazio parametrico, e sia T una statistica consistente per θ (ovvero $T(F_\theta) = \theta\ \forall\theta \in \Theta$). Huber (1981) dimostra che T è asintoticamente efficiente in F_θ se e solo se la funzione di influenza soddisfa

$$IC(x, F_\theta, T) = I(F_\theta)^{-1}(\partial \log f_\theta/\partial\theta),$$

dove f_θ è la funzione di densità di F_θ e $I(F_\theta) = \int(\partial \log f_\theta/\partial\theta)^2 dF_\theta$ è l'informazione di Fisher.

L'efficienza asintotica per un M-estimator si raggiunge utilizzando una funzione ψ del tipo

$$\psi(x) = -c(f_0'((x))(f_0(x))^{-1}$$

che nel noto problema di posizione diventa semplicemente

$$\psi(x - T) = -c(f_0'(x - T))(f_0(x - T))^{-1}.$$

Spesso si usa chiamare la funzione ψ funzione di influenza, poiché quest'utlima ne è fortementre dipendente. ρ, invece, viene chiamata funzione di perdita.

Una volta introdotto il concetto di M-estimator l'estensione di questo strumento ai vari metodi presenti in statistica che si basano sulla massima verosimiglianza è di facile implementazione. Si consideri il caso sulla regressione lineare, $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$. La stima del parametro $\boldsymbol{\beta}$ si ottiene come soluzione del sistema dei minimi quadrati, $\sum(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \min$. Approcciando il problema della stima di $\boldsymbol{\beta}$ per il modello di regressione lineare con gli M-estimator si ottiene il seguente sistema di stima:

$$\sum \rho(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = \min \tag{2.15}$$

oppure (dopo aver opportunamente derivato dalla precedente equazione rispetto a $\boldsymbol{\beta}$)

$$\sum \psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i = 0, \tag{2.16}$$

con $\psi = \partial\rho/\partial\boldsymbol{\beta}$. Se ρ è una funzione convessa utilizzare la (2.15) o la (2.16) è equivalente, altrimenti l'utilizzo dell'equazione (2.16) è sconsigliato (Huber, 1981).

Quando si utilizza uno stimatore di posizione (del tipo $\sum \psi(x_i - T) = 0$), come nel caso della regressione robusta, si deve tenere conto del fatto che in genere lo stimatore T non è invariante rispetto ad un cambiamento di scala:

$$T(bx_1, \ldots, bx_n) \neq bT(x_1, \ldots, x_n),$$

con $b \in \mathbb{R}$. Per risolvere questo inconveniente si ricorre ad una equazione di stima alternativa alla (2.14):

$$\sum \psi\left(\frac{x_i - T}{S}\right) = 0,$$

dove S è uno stimatore invariante rispetto ai cambiamenti di scala, $S(ax_1, \ldots, ax_n) = aS(x_1, \ldots, x_n), \forall a \in \mathbb{R}$. La statistica S si determina risolvendo l'equazione

$$\frac{1}{n}\sum \chi\left(\frac{x_i - \tilde{T}}{S}\right) = b$$

dove $\tilde{T} \neq T$ è uno stimatore ausiliario, tipicamente corrispondente alla mediana delle x. In Huber (1981) si dimostra che T ed S sono consistenti, ovvero $T \to \mu$ e $S \to \sigma$ quasi certamente[5].

Sia ε_i l'errore del modello di regressione lineare relativo all'osservazione i-esima definito da

$$\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Nel caso della regressione robusta l'errore deve essere opportunamente reso invariante rispetto ai cambiamenti di scala (Huber (1981)). La (2.16) assume la forma:

$$\sum \psi\left(\frac{\varepsilon_i}{s}\right)\boldsymbol{x}_i.$$

[5]Nella teoria della probabilità *quasi certamente* indica un evento che si verifica con probabilità 1. Se $(\Omega, \mathcal{A}, P)$ è uno spazio probabilistico, si dice che un evento E in $\mathcal{A}$ si verifica quasi certamente se $P(E) = 1$.

L'utilizzo di una stima di scala uguale per tutte le osservazioni non è vincolante, ad esempio si può ipotizzare un parametro di scala s_i per ogni osservazione. Inoltre se si ritiene che le covariate siano contaminate, si può applicare anche ad esse una funzione d'influenza per renderle robuste, $\sum \psi_1(\frac{\varepsilon_i}{s})\psi_2(\boldsymbol{x}_i)$.

Per la trattazione della teoria asintotica delle stime robuste per il modello di regressione lineare si rimanda a Huber (1981). Per quanto riguarda l'utilizzo pratico degli M-estimator Huber (1981) ammette che l'asimmetria nella distribuzione degli errori causa distorsione nelle stime robuste, anche se in molti casi applicati tale distorsione è trascurabile. Il caso in cui la distorsione può assumere dimensioni importanti si riscontra solo quando le osservazioni sono molto distanti nello spazio dei dati. In generale, esiste un trade-off tra robustezza e correttezza delle stime.

Per un approfondimento sugli M-estimator si consulti, oltre che Huber (1981), anche Andersen (2008), Hoaglin *e altri* (1983) e Wilcox (2003).

2.3.4 M-Quantile

Introduzione

La definizione di M-quantile si deve a Breckling e Chambers (1988). In questo articolo gli autori propongono una soluzione al problema della stima degli estremi di una distribuzione condizionata.

Si è visto nel paragrafo precedente come l'M-estimator sia una stima di un parametro di una distribuzione basata su una funzione di perdita ρ o, equivalentemente, su una funzione d'influenza ψ. Estendendo la definizione di expectile (Newey e Powell, 1987) alla regressione robusta basata sull'M-estimator si ottiene un analogo del quantile, denominato M-quantile, per la distribuzione di interesse (generalmente $Y|X$). Concettualmente l'M-quantile sta alla media di una distribuzione come il quantile sta alla mediana. Nell'ambito della regressione questa tecnica consente di confrontare, restando sullo stesso "piano", il comportamento sulle code della distribuzione di $Y|X$ con il $E[Y|X]$, conservando le proprietà di un M-estimator[6].

I campi di applicazione per questo metodo sono molteplici; soprattutto là dove l'interesse dell'impatto di una covariata sulla variabile di studio riguardi la coda della distribuzione e non solo il "centro". Nel contesto di questa tesi l'interesse sarà rivolto all'uso dell'M-quantile (più precisamente al modello di regressione M-quantile) per ottenere stime per piccole aree (stime della media e dei quantili per piccola area).

Definizione dell'M-quantile

Si consideri il campione univariato $\{x_1, \ldots, x_n\}$ con funzione di ripartizione empirica $F_n(x)$. Sia $q_{1/4}$ il primo quartile del campione $\{x_1, \ldots, x_n\}$. $q_{1/4}$ corrisponde alla mediana della seguente trasformazione:

$$dG_n(x) \propto \begin{cases} \frac{3}{4}dF_n(x) & (x < q_{1/4}) \\ \frac{1}{4}dF_n(x) & \text{altrimenti.} \end{cases}$$

In altri termini $q_{1/4}$ si può ottenere come soluzione di

$$\int \text{sgn}(x - q_{1/4})dG_n(x) = 0,$$

o equivalentemente di

[6] Più precisamente su una qualunque posizione della distribuzione di $Y|X$.

$$\int \psi_{1/4}(x - q_{1/4}) dF_n(x) = 0,$$

dove

$$\psi_{1/4}(t) = \begin{cases} \frac{3}{4}\text{sgn}(t) & (t < 0) \\ \frac{1}{4}\text{sgn}(t) & \text{Altrimenti.} \end{cases}$$

Generalizzando ad un qualsiasi quantile di ordine $\tau \in (0, 1)$ e ad una qualsiasi funzione di influenza $\psi(t)$ si ottiene l'equazione di stima per l'M-quantile di ordine τ (Breckling e Chambers, 1988):

$$\int \psi_\tau(x - \hat{\theta}_\tau) dF_n(x) = 0, \tag{2.17}$$

dove $\hat{\theta}_\tau$ è la stima dell'M-quantile θ_τ e

$$\psi_\tau(t) = \begin{cases} (1 - \tau)\psi(t) & (t < 0) \\ \tau\psi(t) & \text{Altrimenti.} \end{cases} \tag{2.18}$$

Per $\tau = 0.5$ la (2.18) diventa la funzione di influenza per un M-estimator, mentre la soluzione della (2.17) nel caso in cui (nella (2.18)) si assuma $\psi(t) = \text{sgn}(t)$ è il quantile "classico" di ordine τ. In tutti gli altri casi si fa riferimento a θ_τ come al'M-quantile di ordine τ e a $\hat{\theta}_\tau$ come sua stima.

Si noti che, tolto il caso di particolari funzioni ψ, l'M-quantile di ordine τ (o la sua stima) non può essere interpretato come il quantile di ordine τ, cioè non è vero che ci sono τ dati con valore superiore all'M-quantile (di ordine τ) e $1 - \tau$ dati con valore inferiore.

Come gli M-estimator, anche l'M-quantile si può definire tramite una funzione di perdita. Si consideri la funzione di perdita $\rho_\tau(t)$:

$$\rho_\tau(t) = \begin{cases} \rho((1 - \tau)t) & (t < 0) \\ \rho(\tau t) & \text{Altrimenti.} \end{cases}$$

dove $\partial\rho(t)/\partial t = \psi(t)$. La stima dell'M-quantile di ordine τ, $\hat{\theta}_\tau$, si ottiene dal seguente problema di minimo:

$$\int \rho_\tau(x - \theta_\tau) dF_n(x) = \min . \tag{2.19}$$

La soluzione della (2.19) purtroppo non coincide con quella della (2.17), a meno di casi particolari[7]. M. Hanif e K.R.W. Brewer, in un lavoro non pubblicato, dimostrano che l'M-quantile definito tramite la (2.19) approssima meglio il quantile "classico" rispetto alla (2.17). Tuttavia, Breckling e Chambers (1988) sostengono che l'utilizzo della (2.17) sia da preferirsi per definire un analogo del τ-esimo quantile e questo perché **i.** il rapporto $\tau/(1 - \tau)$ è indipendente da ψ, **ii.** per ogni τ e ψ il τ-esimo M-quantile può essere interpretato come l'M-estimator di una distribuzione trasformata la cui media corrisponde al τ-esimo quantile campionario.

[7]Vale il caso in cui $\tau = 0.5$, $\psi(t) = \text{sgn}(t)$ e $\rho(t) = |t|$; vale a dire il caso in cui l'M-quantile corrisponde al quantile tradizionale.

Il modello di regressione M-quantile

Si consideri la variabile di studio Y con valori $\{y_1, \ldots, y_n\}$ e un insieme di vettori di variabili esplicative $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, dove $\boldsymbol{x}_i$, $i = (1, \ldots, n)$, è il vettore delle variabili esplicative associate a y_i, $i = (1, \ldots, n)$. L'obiettivo è quello di modellare la distribuzione condizionata di $Y|X$. E' noto che se $F(Y|X)$ è la funzione di ripartizione della distribuzione di Y condizionata a X e se $Q_y(\tau|X = \boldsymbol{x}) = Q_y(\tau|\boldsymbol{x})$ è il τ-esimo qantile della distribuzione di $Y|X$, allora deve valere $F(Q_y(\tau|\boldsymbol{x})) = \tau$. Nel paragrafo 2.3.1, sulla regressione quantilica, si è ipotizzata una relazione lineare per modellare i quantili della distribuzione di $Y|X$:

$$Q_y(\tau|\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta}(\tau), \tag{2.20}$$

dove $\boldsymbol{X} = \{\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'\}$ è la matrice delle covariate e $\boldsymbol{\beta}(\tau) = \{\beta_1(\tau), \ldots, \beta_n(\tau)\}$ è il vettore degli effetti delle covariate sul quantile di ordine τ della distribuzione di $Y|X$. Anche Breckling e Chambers (1988) propongono un approccio lineare per modellare i quantili sostenendo che nel range di valori delle variabili ausiliarie (x) campionate la (2.20) è una buona approssimazione della relazione reale (sconosciuta) tra i quantili di Y e le covariate. La bontà di tale approssimazione dipende dal metodo usato per stimare il vettore dei coefficienti $\boldsymbol{\beta}(\tau)$, dai dati campionari e dal processo stocastico sottostante ai dati.

Nel paragrafo 2.3.1 si è descritto il metodo proposto da Koenker e Bassett (1978) per la stima dei parametri nella (2.20). Una delle principali critiche mosse a tale metodo è quello di basarsi sulla minimizzazione degli scarti assoluti (2.3)[8]. Breckling e Chambers (1988) sostengono che

> "non ha senso usare il metodo dei minimi quadrati per modellare la media di $Y|X$ e il metodo della minimizzazione degli scarti assoluti per modellarne gli estremi"[9].

Già in risposta a questa critica Newey e Powell (1987) hanno proposto un modello di regressione per l'expectile (si veda in proposito il paragrafo 2.3.2). Breckling e Chambers (1988) propongono una ulteriore generalizzazione del modello di regressione sull'expectile basata sul concetto per cui l'expectile coincidente con la media, invece di corrispondere alla retta di regressione classica, corrisponda alla retta di regressione robusta basata sull'M-estimator.

Si consideri l'equazione di stima del modello proposto per la regressione robusta basato sull'M-estimator:

$$\sum \psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i = 0.$$

La generalizzazione suggerita da Breckling e Chambers (1988) consiste nel sostituire la funzione di influenza ψ con la sua analoga per gli M-quantili ψ_τ. L'M-quantile di ordine τ della distribuzione condizionata di Y dato X è definito come la soluzione $Q_y(\tau, \psi|x)$ dell'equazione di stima

$$\int \psi_\tau(y - Q_y)dF_n(y|x) = 0,$$

da cui, fissati τ e ψ, si ottiene

$$\sum \psi_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i = 0. \tag{2.21}$$

Utilizzando il metodo dei minimi quadrati pesati iterati applicati all'equazione (2.21) si ottiene una stima del vettore dei coefficienti $\boldsymbol{\beta}(\tau)$, dunque:

[8] Opportunamente pesati.

[9] Estremi è inteso come una qualunque posizione nella distribuzione condizionata di $Y|X$.

$$\hat{Q}_y(\tau, \psi|\boldsymbol{x}) = \boldsymbol{X}'\hat{\boldsymbol{\beta}}(\tau),$$

dove $\hat{Q}_y(\tau, \psi|\boldsymbol{x})$ è la stima del τ-esimo M-quantile della distribuzione $Y|X$.

Con questo metodo si modella l'M-quantile τ della distribuzione di $Y|X$ in modo "consistente" rispetto al modello di regressione robusta basato sull'M-estimator.

Nei lavori di Breckling e Chambers (1988), Kokic *e altri* (1997) e Chambers e Tzavidis (2006) viene utilizzata la funzione di influenza *Huber Proposal 2* (Huber, 1964) generalizzata per l'M-quantile. La Huber Proposal 2, ψ_{Huber}, è:

$$\psi_{\text{Huber}}(t) = \begin{cases} -c & (t < -c) \\ t & (-c \leq t \leq c) \\ c & (t > c). \end{cases} \tag{2.22}$$

ψ_{Huber} è una funzione continua e vale $\psi_{\text{Huber}}(0) = 0$ per ogni $c > 0$. Queste caratteristiche assicurano la robustezza qualitativa (Huber, 1981). Generalizzando la funzione di influenza per l'M-quantile si ottiene:

$$\psi_\tau(t) = \begin{cases} \tau\psi_{\text{Huber}}(t) & (t > 0) \\ (1-\tau)\psi_{\text{Huber}}(t) & (\text{Altrimenti}). \end{cases} \tag{2.23}$$

Questa proposta è valida per l'M-quantile in generale e non solo nel caso della regressione M-quantile. Per $c \to \infty$ la definizione di M-quantile coincide con quella di expectile e il modello di regressione M-quantile coincide con il modello di regressione sull'expectile. Posto $\psi_{\text{Huber}}(t) = \psi(t) = t\text{sgn}(t)$ il modello di regressione M-quantile coincide con il modello di regressione quantilica. Dunque il modello di regressione M-quantile è un modello unificato che comprende sia il modello di regressione per l'expectile proposto da Newey e Powell (1987) sia il modello di regressione quantilica.

Nel paragrafo 2.3.3 si è parlato della necessità, nel caso di regressione robusta, di una correzione del residuo con uno stimatore invariante rispetto a cambiamenti di scala. Poiché anche la regressione M-quantilica è una regressione robusta basata sul concetto di M-estimator, è opportuno introdurre uno stimatore di scala adeguato alla funzione di influenza generalizzata (2.23). Chambers e Tzavidis (2006) propongono lo stimatore *MAD (Median Absolute Deviation)*:

$$S = \frac{median|\hat{\varepsilon}_{i\tau}|}{0.6745}, i = 1, \ldots, n,$$

dove $\epsilon_{i\tau} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau)$, con $i = (1, \ldots, n)$ è il residuo di regressione.

Riassumendo, una proposta per la regressione M-quantile $Q_y(\tau|\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta}(\tau)$ è la seguente:

$$\sum_{i=1}^{n} \psi_\tau\left(\frac{\hat{\varepsilon}_{i\tau}}{S}\right)\boldsymbol{x}_i = 0$$

con

$$\psi_\tau\left(\frac{\epsilon_{i\tau}}{S}\right) = \begin{cases} -2(1-\tau)c & (\frac{\epsilon_{i\tau}}{S} < -c) \\ 2(1-\tau)\frac{\epsilon_{i\tau}}{S} & (-c \leq \frac{\epsilon_{i\tau}}{S} < 0) \\ 2\tau\frac{\epsilon_{i\tau}}{S} & (0 \leq \frac{\epsilon_{i\tau}}{S} \leq c) \\ 2\tau c & (\frac{\epsilon_{i\tau}}{S} > c), \end{cases} \tag{2.24}$$

dove $\epsilon_{i\tau} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau)$, $S = \frac{median|\epsilon_{i\tau}|}{0.6745}$, con la funzione Huber 2 che è stata esplicitata direttamente nella (2.24).

Utilizzando il modello di regressione M-quantile non si risolvono tutti i problemi riscontrati nella regressione quantilica. Uno dei problemi maggiori del modello di regressione quantilica è quello del "quantile crossing", ovvero il fatto che due rette riferite a due quantili distinti si intersechino, che si presenta anche per il modello di regressione M-quantile. Breckling e Chambers (1988) considerano sia il caso in cui le rette si intersechino fuori dal range delle osservazioni, sia il caso in cui si intersechino nel range delle osservazioni. Nel primo caso suggeriscono di interpretare il punto dell'intersezione come limite in cui vale la relazione modellata con la regressione M-quantile, nel secondo caso essi ritengono che l'approssimazione lineare dell'M-quantile sia inadeguata. Il problema maggiore legato al quantile crossing deriva dal fatto che in questa condizione non è garantita la non decrescenza di $\hat{Q}_y(\tau|x)$ (Kokic *e altri* (1997)), una proprietà desiderabile per quanto riguarda i quantili. Infatti, non c'è un'estensione per il modello di regressione M-quantile del teorema 2.3.2, dove si garantisce la non decrescenza del quantile stimato in $\bar{\boldsymbol{x}}$ per la regressione quantilica. Anche per l'interpretazione dei coefficienti i problemi sono gli stessi presentati per la regressione quantilica nel paragrafo 2.3.1. Tuttavia, al fine di utilizzare il modello di regressione M-quantile nell'ambito della stima per piccole aree il problema dell'interpretazione dei coefficienti non è rilevante. La differenza rilevante tra regressione M-quantile e quantilica è dovuta al metodo di stima che, nel caso della regressione M-quantile, grazie al metodo dei minimi quadrati pesati iterati e ad una scelta opportuna della funzione di influenza, l'ottimalità nel processo di stima dei parametri è garantita, con l'ulteriore vantaggio di avere stime robuste[10].

Da un punto di vista asintotico la distribuzione asintotica di un M-quantile dipende dalla M-mediana (cioè l'M-quantile definito con $\tau = 0.5$) definita dalla funzione ψ. $\hat{\theta}_\tau$ è una statistica resistente e ha come punto di rottura (*breakdown point*) il minimo tra τ e $1 - \tau$. La robustezza per $\hat{\theta}_\tau$ cresce per $\tau \to 0.5$ quando lo stimatore coincide con l'M-mediana campionaria. Se si utilizza per definire l'M-quantile la (2.23) la varianza asintotica di $\hat{\theta}_\tau$ è pari a

$$A(F) = \int \psi_\tau^2(x) dFx / \delta^2(F)$$

dove $\delta(F) = (1-\tau)\left[F(\theta_\tau) - F(\theta_\tau - c)\right] + \tau\left[F(\theta_\tau + c) - F(\theta_\tau)\right]$. Per τ che tende a 1 oppure a 0 la varianza asintotica è crescente, mentre negli altri casi cresce al crescere di c (parametro della funzione di influenza Huber 2, (2.22)). Per un approfondimento su quanto riportato si consulti Breckling e Chambers (1988). Una nota sul comportamento asintotico dei coefficienti di regressione M-quantilica si trovano in Kokic *e altri* (1997).

2.3.5 Stima per piccole aree con il modello di regressione M-quantile

Introduzione

Come si è visto nel paragrafo 2.2, nell'ambito della stima per piccole aree, nel modello lineare ad effetti misti si assume che la variabilità associata alla distribuzione condizionata di Y dato X può essere spiegata parzialmente da una struttura gerarchica. In questo paragrafo si presenterà un'alternativa proposta da Chambers e Tzavidis (2006) basata sulla regressione M-quantile per modellare la variabilità nella distribuzione di $Y|X$.

[10] Con questo non si vuole dire che le stime robuste siano da preferirsi alle stime non robuste, si sottolinea il fatto che il modello di regressione M-quantile, essendo un modello unificato, può essere specificato in modo da ottenere stime robuste.

La stima della media per piccola area con il modello di regressione M-quantile

Per chiarezza di lettura si riporta la notazione usata nella stima per piccole aree in questo testo. Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita e sia $\boldsymbol{y} = (y_1, \ldots, y_N)'$ il vettore della variabile d'interesse (per tutte le unità della popolazione Ω). Si consideri un campione $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^{m} n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^{m} r_i = N - n$. A livello di popolazione si indicano le aree o i gruppi con Ω_i, in modo che $\Omega = \{\Omega_1, \ldots, \Omega_m\}$. Sia $\boldsymbol{x}_{ji} = \{x_{1,ji}, \ldots, x_{p,ji}\}$ il vettore di p variabili ausiliarie, riferite all'unità j-esima appartenente all'area i, osservabili e note in modo certo e senza errore per tutte le unità della popolazione Ω.

Utilizzando il modello di regressione M-quantilico introdotto nel paragrafo precedente Chambers e Tzavidis (2006) propongono uno metodo alternativo di stima per piccole aree che non dipende da nessuna struttura gerarchica. Kokic *e altri* (1997) e Aragon *e altri* (2006) caratterizzano la variabilità condizionata alle aree della popolazione di interesse con i coefficienti di regressione M-quantile delle singole unità. Per l'unità j-esima, che ha valori y_j per la variabile di studio e $\boldsymbol{x}_j$ per le covariate (escludendo momentaneamente l'indice di area), il coefficiente che identifica l'unità stessa è il valore τ_j tale che $Q_y(\tau_j, \psi|\boldsymbol{x}_j) = y_j$. τ_j è il coefficiente M-quantile che identifica l'unità j-esima a livello di popolazione. Se nella popolazione esiste una struttura gerarchica capace di spiegare parte della variabilità dei dati, allora le unità appartenenti ad uno stesso cluster, definito dalla gerarchia stessa, dovrebbero avere coefficiente M-quantile simile.

Si ipotizzi che per $\tau \in [0, 1]$ l'M-quantile $Q_y(\tau, \psi|\boldsymbol{x})$ segua un modello lineare:

$$Q_y(\tau, \psi|\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta}_\psi(\tau)$$

dove il vettore dei coefficiente di regressione $\boldsymbol{\beta}_\psi(\tau)$, che è funzione di τ, sia tale che per la media di piccola area valga:

$$\bar{Y}_i = N_i^{-1} \left\{ \sum_{j \in s_i} y_j + \sum_{j \in r_i} \boldsymbol{x}_j' \boldsymbol{\beta}_\psi(\tau_j) \right\} \tag{2.25}$$

$$\simeq N_i^{-1} \left\{ \sum_{j \in s_i} y_j + \sum_{j \in r_i} \boldsymbol{x}_j' \boldsymbol{\beta}_\psi(\theta_i) \right\} + N_i^{-1} \sum_{j \in r_i} \boldsymbol{x}_j' \left\{ \frac{\partial \boldsymbol{\beta}_\psi(\theta_i)}{\partial \theta_i} \right\} (\tau_j - \theta_i), \tag{2.26}$$

dove θ_i è la media dei coefficienti M-quantile delle n_i unità j appartenenti all'area i e $\bar{Y}_i$ è la media della piccola area i. θ_i può essere interpretato come coefficiente M-quantile di area. Considerando che il secondo membro della (2.26) è trascurabile, uno stimatore per la media dell'area i è:

$$\hat{\bar{Y}}_i = N_i^{-1} \left\{ \sum_{j \in s_i} y_j + \sum_{j \in r_i} \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \right\}, \tag{2.27}$$

dove $\hat{\cdot}$ rappresenta uno stimatore per una quantità incognita. La definizione del coefficiente M-quantile di area non è univoca e non sono stati proposti criteri di ottimalità per determinarlo. Per esempio, invece di utilizzare la media dei coefficienti M-quantile delle unità appartenenti ad una certa area per

determinare il coefficiente M-quanitle di area, si potrebbe utilizzare la mediana oppure un qualunque criterio di scelta.

Un approccio diverso, ma sempre basato sul modello di regressione M-quantile, alla stima della media di piccola area si basa sull'utilizzo di un coefficiente di regressione di area anziché di un coefficiente M-quantile di area. Dato che l'M-quantile di ordine τ_j rappresenta l'unità j-esima, si potrebbe utilizzare la media (o la mediana) dei coefficienti $\boldsymbol{\beta}_\psi(\tau_j)$ per ottenere un coefficiente di regressione di area. Sostituendo il coefficiente di regressione di area stimato al posto di $\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$ nella (2.27) si otterrebbe uno stimatore alternativo per la media di piccola area. Chambers e Tzavidis (2006) sostengono che se i coefficienti $\boldsymbol{\beta}_\psi(\tau_j)$ sono più simili all'interno dell'area che tra aree, allora i risultati che si ottengono con questo approccio saranno molto simili a quelli che si otterrebbero con l'approccio del "coefficiente M-quanitle di area" (2.27). Inoltre, utilizzando l'approccio del coefficiente M-quanitle di area, si mantiene una struttura "logica" simile a quella del modello lineare ad effetti misti; infatti il coefficiente M-quantile di area può essere interpretato come uno pseudo-effetto casuale di area. Per esempio, se tutti i coefficienti M-quantile di area sono uguali a 0.5, significa che non c'è variabilità tra aree oltre a quella spiegata dalle covariate.

Per ottenere una stima della media per piccola area utilizzando la (2.27) si deve stimare il coefficiente M-quantile di area, θ_i. Si definisca θ_i come la media dei coefficienti M-quantile delle unità dell'area i appartenenti alla popolazione ($\theta_i = N_i^{-1} \sum_{j \in \Omega_i} \tau_j$). Ovviamente non si può presumere di conoscere i dati a livello di popolazione per tutte le variabili. Considerando i soli dati campionari, $j \in s$, una stima del coefficiente M-quantile di area si ottiene semplicemente facendo la media dei coefficienti M-quantile delle unità campionate in una certa area; per l'area i l'M-quantile di area è: $\hat{\theta}_i = \sum_{j \in s_i} \tilde{\tau}_j$, dove $\tilde{\tau}_j$ è il coefficiente M-quantile per le unità campionate[11]. L'M-quantile di una unità campionata è tale che

$$Q_y(\tilde{\tau}_j, \psi | \boldsymbol{x}_j) = \boldsymbol{x}_j' \boldsymbol{\beta}_\psi(\tilde{\tau}_j) = y_j, \tag{2.28}$$

con $j \in s_i$. Per calcolare il coefficiente M-quantile di unità (campionata) si utilizza una griglia molto fitta di punti compresi nell'intervallo $[0, 1]$ e si fa passare per ogni punto di questa griglia una retta di regressione M-quantile, utilizzando i dati campionari per stimare i parametri della retta. Successivamente si identifica per quali punti della griglia vale la (2.28), in questo modo si associa il punto sulla griglia (compreso tra 0 e 1) al coefficiente M-quantile per l'osservazione j. Se la (2.28) non vale per una data osservazione j, allora si ricava il coefficiente M-quantile per quell'osservazione con interpolazione lineare dei punti sulla griglia. Se, per esempio, si ha che $Q_y(a_{\text{greed point}}, \psi | \boldsymbol{x}_j = y_j + \Delta)$ e $Q_y(b_{\text{greed point}}, \psi | \boldsymbol{x}_j = y_j - \Delta)$, allora il coefficiente M-quantile per l'unità j sarà $\tilde{\tau}_j = (a_{\text{greed point}} + b_{\text{greed point}})/2$. Come si è detto precedentemente il coefficiente M-quantile di area può essere calcolato non solo con la media ma anche con la mediana. Inoltre, se ai dati campionari è associato un peso, si può utilizzare tale peso per calcolare la media, o la mediana, dei coefficienti M-quantile di area. L'utilizzo dei pesi campionari in questa fase non garantisce una stima non distorta a livello di disegno campionario.

La stima dei coefficienti di regressione M-quantile si ottiene risolvendo l'equazione (2.21), secondo quanto detto nel paragrafo 2.3.4:

$$\hat{\boldsymbol{\beta}}_\psi(\tau) = (\boldsymbol{X}_s' \boldsymbol{W}_s(\tau) \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s' \boldsymbol{W}_s(\tau) \boldsymbol{y}_s,$$

dove $\boldsymbol{X}_s = (\boldsymbol{x}_{1s}', \ldots, \boldsymbol{x}_{ps}')$ è la matrice nxp delle p covariate campionarie, $\boldsymbol{W}_s(\tau)$ è una matrice diagonale nxn contenente i pesi prodotti nell'ultima iterazione dell'algoritmo dei minimi quadrati

[11] La˜serve solo a distinguere il caso in cui si fa riferimento al coefficiente M-quantile calcolato sulle unità campionarie dal caso del coefficiente M-quantile calcolato sulle unità appartenenti alla popolazione.

pesati iterati utilizzato per calcolare $\hat{\boldsymbol{\beta}}_\psi(\tau)$. Sostituendo a $\hat{\boldsymbol{\beta}}_\psi(\tau)$ nella (2.27) la sua forma esplicita si ottiene:

$$\hat{\bar{Y}}_i = N_i^{-1}\boldsymbol{w}_i'\boldsymbol{y}_s,$$

dove $\boldsymbol{w}_i$ è il vettore dei pesi dell'area i:

$$\boldsymbol{w}_i = 1_{si} + \boldsymbol{W}(\hat{\theta}_i)\boldsymbol{X}_s(\boldsymbol{X}_s'\boldsymbol{W}_s(\hat{\theta}_i)\boldsymbol{X}_s)^{-1}\boldsymbol{t}_{ri}.$$

1_{si} è un vettore di dimensione n con il j-esimo elemento uguale a 1 se l'unità campionata j appartiene all'area i e uguale a 0 altrimenti. $\boldsymbol{t}_{ri}$ è un vettore di dimensione p contenente la somma delle covariate delle unità non campionate nell'area i.

Stima dell'errore quadratico medio dello stimatore della media per piccola area basato sul modello di regressione M-quantile

Chambers e Tzavidis (2006) propongono una stima dell'errore quadratico medio per lo stimatore $\hat{\bar{Y}}_i$. Tale stima si basa sul metodo di stima dell'errore quadratico medio per stimatori lineari corretti robusti (Royall e Cumberland, 1978). Un'approssimazione della varianza di $\hat{\bar{Y}}_i$, calcolata secondo la (2.27), è

$$V(\hat{\bar{Y}}_i - \bar{Y}_i) = N_i^{-2}\left(\sum_{j\in s} u_{ji}^2 V(y_j) + \sum_{j\in r_i} V(y_j)\right), \tag{2.29}$$

dove $u_{ji} = u_i = \boldsymbol{W}(\hat{\theta}_i)\boldsymbol{X}_s(\boldsymbol{X}_s'\boldsymbol{W}_s(\hat{\theta}_i)\boldsymbol{X}_s)^{-1}\boldsymbol{t}_{ri}$. Il problema di questa formulazione è l'interpretazione di $V(y_j)$ nell'approssimazione proposta. In un report non pubblicato (Chambers, 2005) viene proposto un approccio prudenziale nell'interpretazione di $V(y_j)$ che consiste nel non considerare $V(y_j)$ specifico dell'area a cui appartiene l'unità j. Se si utilizza questo approccio, denominato approccio dei residui a livello di popolazione[12], la (2.29) diventa:

$$\begin{aligned} V(\hat{\bar{Y}}_i - \bar{Y}_i) = & \\ & = N_i^{-2}\left(\sum_{j\in s} u_{ji}^2\left(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_\psi(0.5)\right)^2 + \right. \\ & \left. + \sum_{j\in r_i}(N_i - n_i)(n_i-1)^{-1}\sum_{j\in s_i}\left(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_\psi(0.5)\right)^2\right). \end{aligned}$$

In alternativa si può interpretare $V(y_j)$ condizionandosi alle aree; se l'unità j appartiene all'area i la (2.29) diventa:

[12]Denominato in inglese *population-level residuals approach.*

$$V(\hat{\bar{Y}}_i - \bar{Y}_i) =$$
$$= N_i^{-2} \left(\sum_{j \in s} u_{ji}^2 \left(y_j - \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \right)^2 + \right.$$
$$\left. + \sum_{j \in r_i} (N_i - n_i)(n_i - 1)^{-1} \sum_{j \in s_i} \left(y_j - \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \right)^2 \right),$$

ed è denominato approccio dei residui a livello di area[13]. Per i due casi la stima della varianza (2.29) è rispettivamente

$$\hat{V}_i = \sum_{j \in s} \lambda_{ji} \left(y_j - \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(0.5) \right)^2$$

o

$$\hat{V}_i = \sum_{i=1}^{m} \sum_{j \in s_i} \lambda_{ji} \left(y_j - \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \right)^2, \tag{2.30}$$

dove $\lambda_{ji} = N_i^{-2}[u_{ji}^2 + I(j \in i)(N_i - n_i)(n_i - 1)^{-1}]$. Nella prima sommatoria della (2.31) l'indice $i = (1, \ldots, m)$ sottointende che tutte le aree siano state campionate, se così non fosse la sommatoria andrebbe riferita alle sole aree campionate.

Uno stimatore per la distorsione dello stimatore $\hat{\bar{Y}}_i$ è stato proposto da Chambers e Tzavidis (2006). Si supponga che $E[y_j | \boldsymbol{x}_j, j \in i] = \boldsymbol{x}_j' \boldsymbol{\beta}_i$, allora

$$E\left[N_i^{-1} \sum_{j \in s} w_{ji} y_j - \bar{Y}_i \right] \simeq N_i^{-1} \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{ji} \boldsymbol{x}_j' \boldsymbol{\beta}_i - \sum_{j \in i} \boldsymbol{x}_j' \boldsymbol{\beta}_i \right), \tag{2.31}$$

dove $i = (1, \ldots, m)$ sottointende che tutte le aree siano state campionate. Dalla (2.31) si ricava uno stimatore per la distorsione della (2.27):

$$\hat{B}_i = N_i^{-1} \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{ji} \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) - \sum_{j \in i} \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \right).$$

La stima dell'errore quadratico medio dello stimatore (2.27) (media per piccola area) è

$$\hat{MSE}_i = \hat{V}_i + \hat{B}_i^2,$$

dove $\hat{V}_i$ è una delle due versioni proposte precedentemente.

Conclusioni

La stima della media per piccola area con il modello di regressione M-quantile è una nuova alternativa al metodo "tradizionale" di stima per piccole aree basato sul modello di regressione lineare ad effetti

[13] Denominato in inglese *Area-level residuals approach.*

misti. Esso presenta vantaggi e svantaggi rispetto al metodo tradizionale. Innanzitutto il modello M-quantile non è vincolato da assunzioni distribuzionali, al contrario, nel modello lineare ad effetti misti si assume almeno la normalità degli effetti casuali. Secondo, nel modello M-quantile non è necessaria una struttura gerarchica predeterminata per stimare gli effetti di area, al contrario di quanto accade nel modello lineare ad effetti misti. Terzo, il modello M-quantile è un modello robusto verso gli outliers. Tuttavia anche per il modello lineare ad effetti misti è stata sviluppato un metodo robusto (Richardson e Welsh (1996)). Quarto, è facilmente specificabile in modo non parametrico la relazione tra y e $\boldsymbol{x}$ seguendo l'approccio M-quantile, si veda in proposito Aragon *e altri* (2006).

L'aspetto più interessante proposto da Chambers e Tzavidis (2006), per l'argomento trattato in questa tesi, è il predittore a livello di unità per piccola area basato sulla regressione M-quantile:

$$\hat{y}_{ji} = \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i),$$

dove $\hat{y}_{ji}$ è il valore predetto per l'unità j appartenente all'area i con $j \in r_i$. Utilizzando questo predittore, alternativo a quello ottenibile con il modello lineare ad effetti misti, è possibile ottenere una vasta gamma di statistiche per piccole aree che sfruttano tutti i vantaggi esposti precedentemente. Si può ottenere stime di media (presentata in questo paragrafo), totale, quantili, proporzioni e altro, sia utilizzando l'approccio M-quantile sia utilizzando l'approccio tradizionale basato sul modello lineare ad effetti misti. In Chambers e Tzavidis (2006) è stato fatto un confronto per la stima della media tra i due approcci. Nel caso in cui le assunzioni del modello lineare ad effetti misti siano rispettate, sono questi ultimi ad ottenere la performance migliore, in termini di efficienza, negli altri casi invece è preferibile usare l'approccio M-quantile proposto da Chambers e Tzavidis (2006).

Capitolo 3

Stima della Funzione di Ripartizione

3.1 Introduzione

La funzione di ripartizione (*cumulative distribution function* o *cdf* in inglese)[1] di una variabile casuale è una funzione che esprime il modo in cui sono distribuiti ("ripartiti) i valori di una variabili casuale, in forma cumulata. Una definizione più rigorosa è la seguente (Mood *e altri* (1974)):

Definizione 3.1.1 *si dice funzione di ripartizione di una variabile casuale X, indicata da $F(\cdot)$, quella funzione che ha per dominio la retta reale e codominio (o immagine) l'intervallo $[0, 1]$ e che soddisfa $F(y) = P(Y \leq y)$ per ogni numero reale y.*

La conoscenza della funzione di ripartizione implica la conoscenza completa della variabile aleatoria. La stima della funzione di ripartizione è importantissima ai fini dello studio di un certo fenomeno: non ci si limita allo studio di media e varianza di una variabile aleatoria, si vuole avere una conoscenza piena del fenomeno aleatorio oggetto di studio[2]. La funzione di ripartizione di una popolazione finita si indica con $F_N(y)$ mentre se si fa riferimento ad un campione casuale si parla di funzione di ripartizione *empirica* o *campionaria*, e si indica con $F_n(y)$. Si consideri una popolazione finita $\Omega = \{1, \ldots, N\}$ e sia $\{y_1, \ldots, y_N\}$ il valore della variabile casuale Y associata ad ogni unità della popolazione Ω. Si definisce la funzione di ripartizione di una popolazione finita $F_N(y)$, con $y \in (-\infty, +\infty)$, come la proporzione degli elementi della popolazione per cui $y_i \leq y$. Formalizzando il problema $F_N(y)$ è una funzione non decrescente a gradini:

$$F_N(y) = N^{-1}(\#A_y), \tag{3.1}$$

dove A_y indica l'insieme degli elementi della popolazione y_i non maggiore di y, ovvero $A_y = \{i : i \in \Omega, \text{ e } y_i \leq y\}$ e $\#A_y$ indica il numero degli elementi appartenenti all'insieme A_y. Considerando un campione casuale s di n unità estratto da Ω, la funzione di ripartizione empirica (o campionaria) si definisce come la proporzione degli elementi del campione per cui $y_i \leq y$, ovvero:

$$F_n(y) = n^{-1}(\#A_y), \tag{3.2}$$

[1] Anche *distribution function*.

[2] Conoscenza piena del fenomeno aleatorio è da intendersi da un punto di vista probabilistico. Conoscendo la funzione di ripartizione di una variabile casuale è possibile determinare in modo certo la funzione di densità e la funzione generatrice dei momenti, se esiste. Con una di queste tre funzioni il fenomeno aleatorio è noto.

dove il numero degli elementi appartenenti all'insieme A_y è determinato sulla base dei dati campionari. Quando si usa la funzione di ripartizione empirica come stimatore della funzione di ripartizione della popolazione, per rafforzare il suo ruolo di stimatore è consuetudine scrivere $\hat{F}_n(y)$. Nel paragrafo 3.2 saranno presentate le principali proprietà di stimatore della funzione di ripartizione empirica.

Nelle indagini campionarie, spesso, la stima della funzione di ripartizione è uno dei principali obiettivi. Infatti, oltre ad una conoscenza piena del fenomeno oggetto di studio, con la funzione di ripartizione si individuano sottoinsiemi della popolazione che stanno al disopra o al di sotto di un certo valore. Per esempio, per alcuni indicatori di povertà di Laeken, come il *at-risk-of-poverty rate*, si utilizzano i quantili (facilmente ottenibili dalla funzione di ripartizione) della distribuzione del reddito netto equivalente. Inoltre, nel caso in cui la variabile di interesse sia molto asimmetrica o plurimodale la sola conoscenza della media è fuorviante, al contrario con la funzione di ripartizione si identificano facilmente le caratteristiche di una distribuzione come l'asimmetria e la plurimodalità.

La stima della funzione di ripartizione di una certa variabile di studio per una popolazione finita si può ottenere seguendo due diversi approcci: **i.** basato su disegno e **ii.** basato su modello. Il metodo di stima basato su disegno è volto ad ottenere stimatori non distorti (corretti) rispetto al disegno, denominati anche stimatori *design-unbiased*. Nel metodo di stima basato su modello si ipotizza l'esistenza di un modello di superpopolazione per la variabile di studio tramite il quale si cerca di ottenere stimatori non distorti (corretti) rispetto al modello, denominati anche stimatori *model-unbiased*. Nell'approccio basato su modello è necessario conoscere alcune variabili a livello di popolazione, che sono denominate variabili ausiliarie. In letteratura (Rao *e altri*, 1990) sono stati proposti anche stimatori design-unbiased che usano un modello di superpopolazione. Rao *e altri* (1990) hanno proposto uno stimatore della funzione di ripartizione sia model-unbiased sia design-unbiased.

Il problema della stima della funzione di ripartizione nell'ambito della stima per piccole aree è stato principalmente affrontato seguendo l'approccio basato su modello. Per un'applicazione relativa a questo tipo di approccio si veda Tzavidis *e altri* (2008a). Un'applicazione sulla stima della funzione di ripartizione basata su disegno nell'ambito della stima per piccole aree è in corso di pubblicazione (Tzavidis *e altri*, 2008b).

La stima dell'errore quadratico medio di uno stimatore della funzione di ripartizione è un filone di ricerca aperto, soprattutto nell'ambito della stima per piccole aree. In letteratura sono state fatte alcune proposte per la stima dell'errore quadratico medio per uno stimatore della funzione di ripartizione (per esempio Lombardia *e altri* (2003), Chambers *e altri* (1992) e Wu e Sitter (2001)), ma non è stata fatta ancora nessuna proposta nell'ambito della stima per piccole aree. Alcuni approcci sulla stima dell'errore quadratico medio in questione sono legati al modello di superpopolazione utilizzato nella stima della funzione di ripartizione, altri si basano sulle metodologie bootstrap e jackknife. Una proposta di stima dell'errore quadratico medio per uno stimatore della funzione di ripartizione sarà presentata nel capitolo 4.

3.2 La Stima della Funzione di Ripartizione Empirica

La funzione di ripartizione empirica di una popolazione finita per la variabile di studio Y, $F_N(y)$ (3.1), è nota solo nel caso in cui siano noti i valori della variabile di studio per tutte le unità della popolazione. In questo contesto si è interessati ai casi in cui i dati a livello di popolazione per la variabile di studio non sono noti. Quando la variabile di interesse non è nota per tutte le unità della popolazione, si ipotizza di conoscerne il valore per un campione casuale di n unità estratte dalla popolazione stessa. Lo stimatore "candidato" a stimare la funzione di ripartizione (di una popolazione finita) è la funzione di ripartizione empirica.

Si consideri momentaneamente il problema della stima della funzione di ripartizione per una variabile casuale Y. Le principali proprietà statistiche della funzione di ripartizione empirica sono espresse dal seguente teorema:

Teorema 3.2.1 *Sia $Y_1, \ldots, Y_n$ un campione di n variabili casuali distribuite con funzione di ripartizione F, e sia $\hat{F}_n(y)$ la funzione di ripartizione empirica (3.2), allora:*

i. *Per ogni valore y,*

$$E[\hat{F}_n(y)] = F(y) \quad e \quad V[\hat{F}_n(y)] = n^{-1}F(y)(1 - F(y)).$$

Asintoticamente vale $MSE(\hat{F}_n(y)) = V(\hat{F}_n(y)) \to 0$ e quindi $\hat{F}_n(y) \xrightarrow{P} F(y)$.

ii. *(teorema Glivenko-Cantelli)*

$$\sup_y |\hat{F}_n(y) - F(y)| \xrightarrow{q.c.} 0.$$

iii. *(disuguaglianza Dvoretzky-Kiefer-Wolfowitz (DKW))*

$$P\left(\sup_y |F(y) - \hat{F}_n(y)| > \epsilon\right) \leq 2\exp^{-2n\epsilon^2}.$$

Dalla disuguaglianza DKW segue il seguente teorema per l'intervallo di confidenza su $F(y)$:

Teorema 3.2.2 *Sia*

$$\begin{aligned} L(y) &= \max\{\hat{F}_n(y) - \epsilon_n, 0\} \\ U(y) &= \min\{\hat{F}_n(y) + \epsilon_n, 0\}, \end{aligned}$$

dove

$$\epsilon_n = \left((2n)^{-1}\log\frac{2}{\alpha}\right)^{-\frac{1}{2}}.$$

Allora, per ogni F, n e y,

$$P\left(L(y) \leq F(y) \leq U(y)\right) \geq 1 - \alpha.$$

Si prenda in esame la stima della funzione di ripartizione nel caso di popolazioni finite. Si consideri una popolazione (finita) Ω di N unità, e siano $\{y_1, \ldots, y_N\}$ i valori della variabile di interesse misurati sulle N unità di Ω. Si consideri, inoltre, il campione casuale semplice s di $n \leq N$ unità, estratto senza ripetizione da Ω. $s = \{y_1, \ldots, y_n\}$. Anche in questo caso, per n abbastanza grande, per la funzione di ripartizione campionaria valgono le proprietà che sono state presentate[3].

[3] In questo capitolo si considera N noto.

Si consideri il caso in cui il campione s sia estratto secondo un disegno campionario più complesso del campionamento casuale semplice. Sia s un campione selezionato dalla popolazione Ω con un disegno di campionamento $p(\cdot)$ con probabilità di inclusione per l'unità i-esima pari a π_i e probabilità di inclusione dell'unità k-esima e i-esima pari a π_{ki}. Una proposta per la stima della funzione di ripartizione in questo caso è stata fatta da Woodruff (1952). L'idea alla base dello stimatore proposto da Woodruff (1952) è quella di considerare la funzione di ripartizione come la media di una funzione indicatrice:

$$F(y) = N^{-1} \sum_{i \in \Omega} z_{i,y} = \bar{z}_{\Omega,y},$$

dove $z_{i,y}$ è una funzione indicatrice definita per tutta gli elementi della popolazione e per ogni y appartenente ai reali:

$$z_{i,y} = \begin{cases} 1 & (y_i \leq y) \\ 0 & (y_i > y). \end{cases}$$

Una volta definita la funzione di ripartizione come una media di popolazione, si possono utilizzare gli stimatori della media presenti in letteratura per ottenerne una stima. Utilizzando lo stimatore di Horvitz e Thompson (1952) lo stimatore per la funzione di ripartizione è

$$\hat{F}(y) = \hat{\bar{z}}_{s,y} = \left(\sum_{i \in s} z_{i,y} \pi_i^{-1} \right) \left(\sum_{i \in s} \pi_i^{-1} \right)^{-1} = \left(\sum_{i \in s \cap A_y} \pi_i^{-1} \right) \left(\sum_{i \in s} \pi_i^{-1} \right)^{-1}, \qquad (3.3)$$

dove $s \cap A_y$ è l'insieme delle unità campionarie con valori $y_i \leq y$. $\hat{F}(y)$ è una funzione a gradini non decrescente che ha come codominio l'intervallo $[0, 1]$, dunque rispetta quelle proprietà che caratterizzano la funzione di ripartizione.

Uno dei motivi per cui si è interessati alla funzione di ripartizione è il legame che questa ha con i quantili. Come è noto, il quantile q di ordine τ è legato alla funzione di ripartizione secondo la relazione $q = F^{-1}(\tau)$. Una definizione alternativa di quantile è utile nel caso in cui $F(y)$ non sia invertibile: $q = \inf\{y : F(y) \geq \tau\}$. Dalla stima della funzione di ripartizione si può ottenere la stima dei quantili. Per la stima della mediana si ha $\hat{q} = \hat{M} = \hat{F}^{-1}(\tau = 0.5)$.

Una volta ottenuta la stima della funzione di ripartizione o di un quantile si è interessati a trovare un intervallo di confidenza per tale stima. Date due costanti c_1 e c_2 l'intervallo di confidenza per la stima della funzione di ripartizione è definito come

$$Pr(c_1 \leq \hat{F}(q) \leq c_2) = 1 - \alpha.$$

Le costanti c_1 e c_2 si possono determinare in modo approssimato. Ipotizzando che $\hat{F}(q)$ sia approssimativamente normalmente distribuita con media $F(q) = \tau$, dove q è il quantile di ordine τ, le costanti c_1 e c_2 sono determinate da

$$\begin{aligned} c_1 &= \tau - Z_{\alpha/2} V(\hat{F}(q))^{1/2} \\ c_2 &= \tau + Z_{\alpha/2} V(\hat{F}(q))^{1/2}. \end{aligned}$$

La varianza dello stimatore della funzione di ripartizione si ottiene applicando lo stimatore della varianza per lo stimatore della media di tipo Horvitz-Thompson alla (3.3):

$$V(\hat{F}(q)) = V\left(N^{-1}\sum_{i\in s} z_{i,q}\pi_i\right) = N^{-2}\sum\sum_{i\in\Omega}\Delta_{ki}\left((z_{i,q}-\tau)\pi_i\right)\left((z_{k,q}-\tau)\pi_k\right), \tag{3.4}$$

dove $\Delta_{ki} = \pi_{ki} - \pi_i\pi_k$. La (3.4) non è calcolabile poiché q è incognito quindi $z_{i,q}$ è una variabile non osservabile[4]. Per ovviare a questo problema, e ottenere una stima della varianza di $\hat{F}(q)$, è sufficiente sostituire $z_{i,q}$ con $z_{i,\hat{q}}$ (con $i \in s$):

$$\hat{V}(\hat{F}(q)) = N^{-2}\sum\sum_{i\in\Omega}\breve{\Delta}_{ki}\left((z_{i,\hat{q}}-\tau)\pi_i\right)\left((z_{k,\hat{q}}-\tau)\pi_k\right) \tag{3.5}$$

. L'intervallo di confidenza per la stima della funzione di ripartizione è $Pr(c_1 \leq \hat{F}(q) \leq c_2) = 1-\alpha$, dove c_1 e c_2 sono definite utilizzando la (3.5):

$$c_1 = \tau - Z_{\alpha/2}V(\hat{F}(q))^{1/2} \tag{3.6}$$

$$c_2 = \tau + Z_{\alpha/2}V(\hat{F}(q))^{1/2}. \tag{3.7}$$

Per ottenere un intervallo di confidenza per il quantile q si consideri la seguente relazione:

$$Pr(c_1 \leq \hat{F}(q) \leq c_2) = Pr(\hat{F}^{-1}(c_1) \leq q \leq \hat{F}^{-1}(c_2)). \tag{3.8}$$

Dalla (3.8) e dalla (3.6) l'intervallo di confidenza per il quantile q di ordine τ è

$$[\hat{F}^{-1}(\tau - Z_{\alpha/2}\hat{V}(\hat{F}(q))^{1/2}), \hat{F}^{-1}(\tau + Z_{\alpha/2}\hat{V}(\hat{F}(q))^{1/2})]. \tag{3.9}$$

Essendo $\hat{q}$ una stima consistente per q allora $\hat{V}(\hat{F}(q))$ è una stima consistente per $V(\hat{F}(q))$ e ciò giustifica la procedura che è stata seguita. In Särndal *e altri* (1992) si avverte di utilizzare questo metodo con molta cautela e di non utilizzarlo per niente nel caso di piccoli campioni.

Lo stimatore della funzione di ripartizione presentato non è generalmente adeguato nel caso della stima per piccole aree. Infatti, anche se lo stimatore della funzione di ripartizione è accettabile, la stima dell'errore quadratico medio (che in questo caso coincide con la stima della varianza) non è applicabile per campioni di piccole dimensioni.

3.2.1 La Funzione di Ripartizione Empirica nell'Ambito della Stima per Piccole Aree: una Simulazione

Per verificare il comportamento della stima della funzione di ripartizione nell'ambito della stima per piccole aree è stata fatta una simulazione Monte Carlo. La simulazione si snoda su due punti principali: **i.** una simulazione basata su una popolazione generata tramite un modello ,e **ii.** una simulazione basata su un dataset esistente.

Nel primo caso sono state generate due popolazioni, Ω_1 e Ω_2, divise per aree. A entrambe le popolazioni sono state associate due variabili, Y e X. La variabile Y è stata generata con un modello lineare ad effetti misti del tipo

$$y_{ji} = 1 + x_{ji} + u_i + \varepsilon_{ji},$$

[4]Per ricedere questi concetti si consulti Särndal *e altri* (1992).

dove y_{ji} e x_{ji} identificano rispettivamente la variabile di studio e la variabile ausiliaria per l'unità j-esima appartenente all'area i-esima con $j = (1, \ldots, N_i)$ e $i = (1, \ldots, m)$; N_i indica la dimensione della popolazione nell'area i. Sia per Ω_1, sia per Ω_2 m è uguale a 30 e N_i varia tra 50 e 450. m e N_i sono mantenuti fissi in ogni iterazione della simulazione Monte Carlo. La dimensione di Ω_1 e Ω_2 è pari a $N = 7430$ unità. In Ω_1, x_{ji} è stata generata da una distribuzione normale con media μ_i e varianza 1. La media μ_i, fissa nella simulazione, varia tra 20 e 100. u_i è stato generato da una distribuzione normale con media 0 e varianza 1, mentre ε_{ji} da una distribuzione normale con media 0 e varianza 16. In Ω_2, x_{ji} è stata generata da una distribuzione χ^2 con gradi di libertà variabili tra 2 e 5 e tenuti fissi nella simulazione. u_i è stato generato centrando una distribuzione χ^2 con 1 grado di libertà mentre ε_{ji} centrando una distribuzione χ^2 con 3 gradi di libertà. Le due popolazioni sono state generate in questo modo per studiare lo stimatore della funzione di ripartizione proposto sia in una situazione di simmetria dei dati, nel caso di Ω_1, sia in una situazione asimmetrica, come nel caso di Ω_2. Le popolazioni sono generate secondo quanto detto ad ogni iterazione Monte Carlo. Dalle popolazioni Ω_1 e Ω_2 è stato estratto un campione casuale semplice s. Da ogni area sono stati estratte n_i unità, con $n_i = N_i/10$. Il campione s ha dimensione totale $n = 743$ unità. Tramite i dati campionari è stata stimata ad ogni iterazione Monte Carlo la funzione di ripartizione per piccola area ed è stato costruito su tale stima un intervallo di confidenza al 95%. E' stato calcolato l'errore relativo per la stima dei percentili 25, 50 e 75 e il tasso effettivo di copertura. L'errore relativo per un dato percentile, τ, è stato calcolato come segue:

$$RB(\tau, i) = H^{-1} \sum_{h=1}^{H} \frac{\hat{F}_{i,h}^{-1}(\tau) - F_{i,h}^{-1}(\tau)}{F_{i,h}^{-1}(\tau)}, \tag{3.10}$$

dove $F_{i,h}^{-1}(\tau)$ è il vero valore del quantile di ordine τ nell'area i nell'iterazione h della simulazione, $\hat{F}_{i,h}^{-1}(\tau)$ è la stima del quantile di ordine τ nell'area i nell'iterazione h e H è il numero di simulazioni Monte Carlo effettuate. Nella simulazione presentata H è uguale a 500. Il tasso effettivo di copertura è stato ottenuto calcolando la percentuale di volte in cui il vero valore è caduto all'interno dell'intervallo di confidenza:

$$CR(q, i) = H^{-1} \sum_{h=1}^{H} I(F_{i,h}^{-1}(\tau) \in CI(\tau, i, h)), \tag{3.11}$$

dove $CI(\tau, i, h)$ è l'intervallo di confidenza per il quantile di ordine τ nell'area i nell'iterazione h, calcolato secondo la (3.9) ed I è la funzione indicatrice. Nella tabella 3.1 sono stati riportati l'errore relativo in percentuale e il tasso di copertura. Vista la quantità di aree elevata (30) si è preferito riportare la distribuzione degli indici RB (3.10) e CR (3.11) tra le aree.

Dalla tabella 3.1 si nota che lo stimatore usato è molto preciso, infatti l'errore relativo medio non supera mai l'1%. Anche il livello di copertura si presenta sempre distribuito intorno al valore nominale del 95%. Effettivamente questo è un risultato inatteso. Infatti, secondo quanto suggerito da Särndal *e altri* (1992), non è prudente usare l'intervallo di confidenza calcolato con la (3.9) per piccoli campioni in quanto non è garantita la copertura. Nel nostro esperimento, nella maggior parte dei casi, la copertura ottenuta è stata molto vicina a quella nominale del 95%, sia nel caso della popolazione Ω_1, con variabile ausiliaria ed errori normali, sia nel caso della popolazione Ω_2, con variabile ausiliaria ed errori χ^2. Questo fatto è probabilmente dovuto alle condizioni "ideali in cui si è svolta la simulaizone (mancanza di outliers, dati generati da modello con disturbi non molto grandi, distribuzioni unimodali, ecc.). E' utile osservare il comportamento dello stimatore proposto con dati reali.

Tabella 3.1: Distribuzione tra le aree dell'errore relativo e del tasso di copertura dello stimatore empirico per la funzione di ripartizione.

		Ω_1		Ω_2	
		RB %	CR	RB %	CR
$\tau = 0.25$	Min	-0.007	0.88	0.040	0.86
	$Qu.1$ quart.	0.086	0.94	0.092	0.93
	Mediana	0.133	0.95	0.158	0.94
	Media	0.177	0.95	0.203	0.94
	$Qu.3$ quart.	0.231	0.96	0.287	0.95
	Max	0.566	0.97	0.705	0.97
$\tau = 0.50$	Min	-0.173	0.93	-0.035	0.92
	$Qu.1$ quart.	-0.056	0.94	0.020	0.94
	Mediana	-0.010	0.94	0.051	0.95
	Media	-0.015	0.95	0.057	0.95
	$Qu.3$ quart.	0.030	0.95	0.088	0.95
	Max	0.166	0.97	0.152	0.96
$\tau = 0.75$	Min	-0.578	0.86	-0.899	0.85
	$Qu.1$ quart.	-0.260	0.93	-0.209	0.94
	Mediana	-0.130	0.94	-0.125	0.95
	Media	-0.186	0.94	-0.163	0.94
	$Qu.3$ quart.	-0.065	0.95	-0.052	0.96
	Max	-0.029	0.97	0.010	0.97

La seconda simulazione è stata fatta utilizzando il campione Leaving Standard Measurement Survey Albania (2003), d'ora in avanti campione Albania. Vista la difficoltà di generare artificialmente una popolazione realistica, si è utilizzato il campione Albania come popolazione da cui estrarre un campione su cui calcolare stima e intervallo di confidenza della funzione di ripartizione per piccola area. Il campione Albania, la nostra popolazione, è composto da 3591 unità divise in 36 aree. In ogni area è stato estratto un campione casuale semplice di dimensione pari a 1/10 della dimensione della popolazione dell'area stessa; nelle aree in cui il numero di unità è risultato inferiore a 50 è stato estratto un campione di 5 unità. Il campione finale è risultato di 398 unità. Nelle 500 simulazioni Monte Carlo la popolazione è rimasta invariata mentre un campione è stato estratto ad ogni iterazione. Su ogni campione estratto sono stati stimati il 25-esimo, il 50-esimo e il 75-esimo percentile del reddito equivalente ed il rispettivo intervallo di confidenza ad un livello di fiducia del 95%. Nella tabella 3.2 sono stati riportati l'errore relativo in percentuale e il tasso di copertura. Vista la quantità di aree elevata (36), si è preferito riportare la distribuzione degli indici RB (3.10) e CR (3.11) tra le aree.

Dai dati riportati in tabella 3.2 si nota chiaramente come il livello di copertura sia lontano dal livello nominale del 95% per la stima dei quantili 0.25 e 0.75. Nel caso del quantile 0.50 la copertura è leggermente al di sotto del livello nominale. L'errore relativo è molto contenuto, fatta eccezione per alcune aree in cui raggiunge anche il 30%. In conclusione, non è indicato utilizzare come stimatore della funzione di ripartizione la funzione di ripartizione empirica nel caso delle piccole aree. Il rischio principale è quello di una errata inferenza che porta ad una sottocopertura nell'intervallo di confidenza per la stima dei quantili, in modo particolare allontanandosi dal 50-esimo percentile.

Tabella 3.2: Distribuzione tra le aree dell'errore relativo e del tasso di copertura dello stimatore empirico per la funzione di ripartizione.

		RB %	CR
$\tau = 0.25$	Min	-0.686	0.72
	$Qu.1$ quart.	2.744	0.77
	Mediana	4.630	0.81
	Media	7.373	0.85
	$Qu.3$ quart.	9.648	0.95
	Max	30.130	0.97
$\tau = 0.50$	Min	-7.734	0.72
	$Qu.1$ quart.	-0.123	0.92
	Mediana	1.194	0.93
	Media	2.125	0.93
	$Qu.3$ quart.	4.175	0.94
	Max	11.780	0.96
$\tau = 0.75$	Min	-13.090	0.72
	$Qu.1$ quart.	-5.577	0.77
	Mediana	-2.648	0.84
	Media	-3.155	0.85
	$Qu.3$ quart.	-1.001	0.94
	Max	5.532	0.97

Queste simulazioni non vogliono essere esaustive sull'utilizzo della funzione di ripartizione empirica nell'ambito della stima per piccole aree. Lo scopo è stato quello di verificare il comportamento dello stimatore funzione di ripartizione empirica per piccole aree in casi realistici. Tuttavia non sono stati considerati casi di campioni estratti con un disegno di campionamento complesso oppure dataset con particolari peculiarità. Dall'analisi fatta risulta comunque la necessità di uno stimatore alternativo per la funzione di ripartizione (e quindi dei quantili) nell'ambito della stima per piccole aree.

La spinta verso uno stimatore alternativo alla funzione di ripartizione empirica è derivata, anche, dal fatto che quest'ultima non utilizza informazioni ausiliarie. Nel paragrafo successivo saranno introdotti due metodi per la stima della funzione di ripartizione che si avvalgono di variabili ausiliarie nel caso siano disponibili a livello di popolazione.

3.3 La Stima della Funzione di Ripartizione con l'Utilizzo di Variabili Ausiliarie

3.3.1 Introduzione

I principali approcci alla stima della funzione di ripartizione con l'utilizzo di variabili ausiliarie sono due: **i.** quello basato su modello e **ii.** quello basato si disegno.

In questo lavoro l'attenzione è rivolta principalmente sul metodo di stima proposto da Chambers e Dunstan (1986), d'ora in avanti denominato stimatore CD, che rientra nella tipologia degli stimatori basati su modello. In questo paragrafo sarà presentata una versione per piccola area dello stimatore

CD, in cui saranno utilizzati il modello lineare ad effetti misti ed il modello di regressione M-quantile (introdotti nel capitolo 2).

Per gli stimatori basati su disegno sarà introdotto lo stimatore proposto da Rao *e altri* (1990), d'ora in avanti denominato stimatore RKM, anche nella versione adattata alla stima per piccole aree. Per una trattazione esaustiva dello stimatore RKM si consulti Rao *e altri* (1990) e Wu e Sitter (2001); si consulti Tzavidis *e altri* (2008b) per un primo approccio sull'utilizzo dello stimatore RKM nell'ambito della stima per piccole aree.

3.3.2 Lo Stimatore Chambers-Dunstan

Nella stima basata su modello si ipotizza che un certo carattere della popolazione (finita) sia generato in funzione di un insieme di variabili ausiliarie e di un certo errore.

Si consideri il carattere osservabile Y di una popolazione finita Ω di N unità a cui si associa la variabile di interesse Y, $y_1, \ldots, y_N$. Ad ogni elemento della popolazione si associ un vettore di p variabili ausiliarie osservabili, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, dove $\boldsymbol{x}_i = x_1, \ldots, x_p$ con $i = 1, \ldots, N$. Si ottengono N coppie $(y_i, \boldsymbol{x}_i)$ con $i = 1, \ldots, N$. Si ipotizzi che esista una relazione tra il carattere Y e le variabili ausiliarie $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ tale per cui

$$y_i = f(\boldsymbol{x}_i; \varepsilon_i), \tag{3.12}$$

dove ε_i è una variabile casuale non osservabile con determinate proprietà statistiche. Se la relazione (3.12) è vera, il carattere Y è completamente determinato dalle variabili ausiliarie e dal termine di errore ε. In questo caso Ω è denominata superpopolazione ed Y è intesa come variabile casuale.

Di solito si ipotizza che il termine di errore ε_i per l'unità i sia generato da una variabile casuale normalmente distribuita e non osservabile. Una forma funzionale che si può utilizzare nel caso in cui il carattere Y sia continuo è quella della regressione lineare:

$$y_i = \alpha + \boldsymbol{\beta}\boldsymbol{x}_i + \varepsilon_i, \tag{3.13}$$

dove α è una costante, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ è il vettore dei coefficienti di regressione e ε_i è un errore normalmente distribuito con media 0 e varianza σ_i^2.

La relazione che esiste tra variabile di interesse e variabili ausiliarie si utilizza in questo lavoro nel caso in cui la variabile di interesse non sia nota per tutte le unità della popolazione ma solo per un suo sottoinsieme. Si ipotizzi di estrarre un campione casuale semplice s di dimensione $n \leq N$ dalla popolazione Ω, e di osservare il carattere Y per le unità appartenenti ad s ($s \subset \Omega$). Ipotizzando che per il carattere Y valga la relazione (3.13), utilizzando il metodo dei mini quadrati ordinari si ottiene una stima ottima lineare e corretta dei parametri α e $\boldsymbol{\beta}$. Se il valore delle variabili ausiliarie è noto anche per le unità non appartenenti al campione, utilizzando la stima dei parametri α e β si ottiene una stima del carattere Y per le unità non campionate:

$$\hat{y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}\boldsymbol{x}_i \quad i \notin s, \quad i \in \Omega - s$$

dove $\hat{y}_i$ è il valore predetto per y_i con $i \in \Omega - s$, $\hat{\alpha}$ e $\hat{\boldsymbol{\beta}}$ sono le stime dei coefficienti di regressione e $\boldsymbol{x}_i$ è il vettore di variabili ausiliarie per l'unità i-esima ed è noto per $i \in \Omega - s$. Per un approfondimento sulle proprietà statistiche di $\hat{y}$ si consulti Green (2000), mentre per un approfondimento sul modello di superpopolazione si consulti Särndal *e altri* (1992).

Per stimare la funzione di ripartizione di una popolazione finita utilizzando lo stimatore CD è necessario supporre che la variabile di studio Y sia generata da un modello di superpopolazione.

Si consideri la popolazione finita Ω formata da N unità, $i = 1, \ldots, N$. Si consideri i caratteri Y e X della popolazione Ω, con Y carattere continuo. Per le N unità della popolazione si ha la coppia (y_i, x_i), $i = 1, \ldots, N$. Si ipotizzi di conoscere in modo certo e senza errore il carattere X per tutte le unità della popolazione e di non conoscere il carattere Y. Si estragga un campione s da Ω di $n \leq N$ unità e su ogni unità si misuri il carattere Y. Condizionandoci al campione estratto, il carattere Y è noto per le n unità appartenenti al campione s. Si consideri il seguente modello di superpopolazione:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{3.14}$$

dove α e β sono i coefficienti di regressione e gli ε_i sono variabili casuali normali indipendenti e identicamente distribuite con media 0 e varianza costante σ^2.

L'obiettivo è quello di stimare la funzione di ripartizione del carattere Y per la popolazione Ω:

$$F_N(t) = N^{-1} \sum_{i \in \Omega} I(y_i \leq t). \tag{3.15}$$

La (3.15) è un'espressione equivalente alla (3.1) in cui si utilizza la funzione indicatrice I invece dell'operatore $\#$ e dell'insieme A_y[5].

L'idea alla base dello stimatore CD è quella di scomporre la (3.15) in due parti, una riferita alle unità campionate e una alle unità non campionate:

$$F_N(t) = N^{-1} \left[\sum_{i \in s} I(y_i \leq t) + \sum_{k \in r} I(y_k \leq t) \right], \tag{3.16}$$

dove r indica l'insieme delle $N - n$ unità non campionate, $r = \Omega - s$. Nella (3.16) la parte non nota riguarda solo il termine $\sum_{k \in r} I(y_k \leq t)$ che deve essere stimato. Intuitivamente si potrebbe pensare di stimare la (3.16) utilizzando un predittore lineare per le unità non campionate, concorde con il modello di superpopolazione ipotizzato. In questo modo lo stimatore della funzione di ripartizione, denominato stimatore *naïve*, avrebbe la seguente forma:

$$\tilde{F}_N(t) = N^{-1} \left[\sum_{i \in s} I(y_i \leq t) + \sum_{k \in r} I(\hat{y}_k \leq t) \right], \tag{3.17}$$

dove $\hat{y}_k = \hat{\alpha} + x_k \hat{\beta}$, con $\hat{\alpha}$ e $\hat{\beta}$ stime ottenute con il metodo dei minimi quadrati ordinari. Secondo Chambers e Dunstan (1986) lo stimatore $\tilde{F}_N(t)$ non è asintoticamente corretto nel caso in cui l'errore nel modello di superpopolazione sia eteroschedastico o asimmetrico. Intuitivamente lo stimatore $\tilde{F}_N(t)$ sarebbe corretto se $\sum_{k \in r} I(\hat{y}_k \leq t)$ fosse uno stimatore corretto di $\sum_{k \in r} I(y_k \leq t)$. Si consideri il termine $\sum_{k \in r} I(\hat{y}_k \leq t)$, esso può essere riscritto come

$$\sum_{k \in r} I(\hat{y}_k \leq t) = \sum_{k \in r} I(y_k - y_k + \hat{y}_k \leq t) = \sum_{k \in r} I(y_k + (\hat{y}_k - y_k) \leq t) =$$
$$= \sum_{k \in r} I(y_k + \varepsilon_k \leq t) \neq \sum_{k \in r} I(y_k \leq t).$$

Se l'errore ε_k è distribuito simmetricamente intorno a 0 allora lo stimatore $\tilde{F}_N(t)$ è corretto rispetto al modello. Tzavidis *e altri* (2008b) notano che l'assunzione di simmetria intono allo 0 dell'errore

[5] $I(u \leq t) = \begin{cases} 1 & (u \leq t) \\ 0 & (altrimenti). \end{cases}$

ε è accettabile per valori di t nella $\sum_{k\in r} I(y_k \leq t)$ vicini alla mediana, mentre non è accettabile altrimenti. In pratica lo stimatore naïve della funzione di ripartizione non dovrebbe essere usato se non per stimare la mediana della variabile di interesse Y.

Chambers e Dunstan (1986) propongono di utilizzare una trasformazione $T(y, \lambda)$, dipendente dal parametro incognito λ. Si consideri la variabile casuale $W = T(y, \lambda)$ con $T(y, \lambda)$ tale che W sia indipendente e identicamente distribuita. Sia G la funzione di ripartizione di W e sia G_n la stima di G ottenuta con la funzione di ripartizione campionaria. Sia λ_n lo stimatore campionario di λ. Uno stimatore per $\sum_{k\in r} I(y_k \leq t)$ è $\sum_{k\in r} G_n\left(T(t, \lambda_n)\right)$. Sostituendo tale stimatore nella (3.16) si ottiene lo stimatore CD (Chambers e Dunstan, 1986) per la funzione di ripartizione di una popolazione finita:

$$\hat{F}_N(t) = N^{-1}\left[\sum_{i\in s} I(y_i \leq t) + \sum_{k\in r} G_n\left(T(t, \lambda_n)\right)\right]. \tag{3.18}$$

Adottando il modello di superpopolazione ipotizzato nella (3.14), la trasformazione T è $T(y, \lambda = (\alpha, \beta)) = W_i = y_i - \alpha - \beta x_i$. Considerando le ipotesi fatte sul modello (3.14), la stima ottima lineare e corretta per i parametri α e β si ottiene con il noto metodo dei minimi quadrati ordinari. Il residuo di regressione ϵ_i è definito come

$$\epsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i, \quad i \in s \tag{3.19}$$

dove $\hat{\alpha}$ e $\hat{\beta}$ sono le stime dei coefficienti di regressione ottenute con il metodo dei minimi quadrati ordinari. La funzione di ripartizione G della variabile casuale W è stimata in questo caso con $\hat{G}_n(t) = n^{-1}\sum_{i\in s} I(\epsilon_i \leq t)$, ovvero con la funzione di ripartizione campionaria[6]. Sostituendo $T(y, \lambda)$ con ϵ_i e $G(t)$ con $\hat{G}_n(t)$ si ottiene lo stimatore CD della funzione di ripartizione nel caso di modello lineare semplice di superpopolazione:

$$\hat{F}_N(t) = N^{-1}\left[\sum_{i\in s} I(y_i \leq t) + \sum_{k\in r} \hat{G}_n(t - \hat{\alpha} - \hat{\beta} x_k)\right]. \tag{3.20}$$

La (3.20) può essere riscritta nella forma seguente:

$$\hat{F}_N(t) = N^{-1}\left[\sum_{i\in s} I(y_i \leq t) + \sum_{k\in r} n^{-1}\sum_{i\in s} I(\hat{\alpha} + \hat{\beta} x_k + \epsilon_i \leq t)\right],$$

dove $\hat{\alpha} + \hat{\beta} x_k + \epsilon_i$, con $k \in r$ e $i \in s$, è un predittore per la variabile di studio delle unità non osservate.

Chambers *e altri* (1992) dimostrano che lo stimatore $\hat{F}_N(t)$ (3.20) è asintoticamente consistente e corretto rispetto al modello lineare di superpopolazione (3.14). Si consideri le seguenti assunzioni:

C-i. La densità $g(\cdot) = dG(\cdot)/d\cdot$ è una funzione delimitata ed esistono derivata prima e seconda.

C-ii. Per n ed N che tendono ad infinito allora n/N tende a π, dove $\pi \in [0, 1]$.

C-iii. $\{x_1, \ldots, x_N\}$, i valori della variabile ausiliaria X per tutta la popolazione, appartengono ad un intervallo finito. La variabile ausiliaria per le unità campionate, $x_i, i \in s$, e per le unità non campionate, $x_k, k \in r$, sono generate dalla stessa funzione di densità δ e vale l'approssimazione

[6] Il simbolo $\hat{\cdot}$ è utilizzato per sottolineare la funzione di stimatore che ha in questo caso la funzione di ripartizione campionaria.

$$n^{-1}\sum_{i\in s} I(x_i \leq t) \rightarrow \int_{-\infty}^{t} \delta(u)du, \quad (N-n)^{-1}\sum k \in rI(x_k \leq t) \rightarrow \int_{-\infty}^{t} \delta(u)du.$$

Chambers *e altri* (1992) provano che se valgono le condizioni **C-i**,**C-ii** e **C-iii** allora

$$E[\hat{F}_N(t) - F_n(t)] = O(n^{-1}),$$

ovvero lo stimatore CD ($\hat{F}_N(t)$) è asintoticamente corretto rispetto al modello (3.14) e

$$\hat{F}_N(t) \rightarrow \int_{-\infty}^{+\infty} G(t-\alpha-\beta x)\delta(x)dx,$$

lo stimatore CD ($\hat{F}_N(t)$) è asintoticamente consistente. Correttezza e consistenza asintotica (sempre rispetto al modello) sono dimostrate solo nel caso in cui il modello di superpopolazione sia quello lineare semplice (3.14), ma, come suggerito in Chambers *e altri* (1992), è facile estendere tali dimostrazione per altri modelli.

La stima dell'errore quadratico medio dello stimatore CD è stata affrontata in Chambers *e altri* (1992), dove viene proposta una versione analitica della varianza asintotica per lo stimatore CD, sempre nel caso in cui il modello per la variabile di interesse, Y, sia quello lineare semplice (3.14). Si consideri le seguenti definizioni:

i. $g(\varepsilon) = dG(\varepsilon)/d\varepsilon$, dove $g(\varepsilon)$ è la densità dell'errore del modello (3.14) e $G(\varepsilon)$ ne è la funzione di ripartizione.

ii. $\sigma^2 = \int \varepsilon^2 g(\varepsilon)d\varepsilon$, la varianza dell'errore nel modello di regressione (3.14).

iii. $\mu = \int x\delta(x)dx$, la media della variabile ausiliaria.

iv. $\varsigma^2 = \left(\int x^2\delta(x)dx\right) - \mu^2$, la varianza asintotica della variabile ausiliaria.

v. $(x \wedge y) = \min(x, y)$.

vi. $\pi = n/N$.

La varianza asintotica dello stimatore CD espresso nella (3.20) (sotto ipotesi di modello lineare semplice per la Y) proposta da Chambers *e altri* (1992) è

$$
\begin{aligned}
V[\hat{F}_N(t) - F_N(t)] = \\
&= n^{-1}(1-\pi)^2 \left\{ \varsigma^{-2}\sigma^2 \left(\int (x-\mu) g(t-\alpha-\beta x)\delta(x)dx \right)^2 + \right. \\
&+ \iint G[(t-\alpha-\beta x) \wedge (t-\alpha-\beta y)]\delta(x)\delta(y)dxdy - \\
&\left. - \left(\int G(t-\alpha-\beta x)\delta(x)dx \right)^2 \right\} + \\
&+ N^{-1}(1-\pi) \int [G(t-\alpha-\beta x) - G(t-\alpha-\beta x)^2]\delta(x)dx + o(n^-1).
\end{aligned}
\tag{3.21}
$$

La (3.21) si basa sui parametri non osservabili e incogniti α, β, σ e ς, e sulle funzioni di densità incognite δ e g, quindi deve essere stimata. Una stima della (3.21) è stata proposta da Wang e Dorfman (1996). Sia

$$\hat{G}_n(t) = n^{-1} \sum_{i \in s} I(\epsilon_i \leq t),$$

la stima della funzione di ripartizione G dell'errore di regressione, ε, nel modello di superpopolazione, ottenuta dalla funzione di ripartizione empirica del residuo di regressione ϵ (3.19), e sia

$$\hat{g}(t) = (nh)^{-1} \sum_{i \in s} K((\epsilon - t)h^{-1}),$$

la stima kernel della funzione di densità g dell'errore di regressione, ε, nel modello di superpopolazione, ottenuta con una bandwidth h e un kernel $K(\cdot)$[7]. Sia inoltre $\hat{\sigma^2} = (n-2)^{-1} \sum_{i \in s} \epsilon_i^2$, la stima per la varianza dell'errore di regressione (nel modello di superpopolazione) e $\hat{\sigma_x^2} = (n-1)^{-1} \sum_{i \in s} (x_i - \bar{x})^2$, la stima della varianza della variabile ausiliaria x (dove $\bar{x}$ è la media campionaria della variabile ausiliaria x). La stima per la (3.21) proposta da Wang e Dorfman (1996) è:

$$
\begin{aligned}
\hat{V}[\hat{F}_N(t) - F_N(t)] = \\
&= n^{-1}(1-\pi)^2 \left\{ \hat{\sigma_x}^{-2}\hat{\sigma}^2 \left((N-n)^{-1} \sum_{k \in r} (x_k - \bar{x}) \hat{g}(t-\hat{\alpha}-\hat{\beta}x_k) \right)^2 + \right. \\
&+ (N-n)^{-1} \sum_{i \in s} \sum_{k \in r} \hat{G}_n[(t-\hat{\alpha}-\hat{\beta}x_i) \wedge (t-\hat{\alpha}-\hat{\beta}x_k)] - \\
&\left. - \left((N-n)^{-1} \sum_{k \in r} \hat{G}_n(t-\hat{\alpha}-\hat{\beta}x_k) \right)^2 \right\} + \\
&+ N^{-1}(1-\pi)(N-n)^{-1} \sum_{k \in r} [\hat{G}_n(t-\hat{\alpha}-\hat{\beta}x_k) - \hat{G}_n(t-\hat{\alpha}-\hat{\beta}x_k)^2].
\end{aligned}
\tag{3.22}
$$

[7]La funzione kernel serve a determinare il tipo di smooth in una stima non parametrica mentre la bandwidth determina l'entità dello smooth, per un approfondimento si consulti Wasserman (2004).

Wu e Sitter (2001) sostengono che lo stimatore per la varianza proposto da Wang e Dorfman (1996) è instabile, nel senso che non è garantita la non negatività della varianza stimata. Per risolvere questo problema Wu e Sitter (2001) propongono uno stimatore più robusto e che garantisce la non negatività della stima, una proprietà certamente desiderabile per uno stimatore della varianza. Indicando con $x \vee y$ il maggiore dei termini tra x e y si può riscrivere la varianza (3.21) dello stimatore CD nel caso di modello lineare semplice di superpopolazione (3.14) come:

$$
\begin{aligned}
&V[\hat{F}_N(t) - F_N(t)] = \\
&= n^{-1}(1-\pi)^2 \left\{ \varsigma^{-2}\sigma^2 \left(\int (x-\mu) g(t-\alpha-\beta x)\delta(x)dx \right)^2 + \right. \\
&\left. + \iint G[(t-\alpha-\beta x) \wedge (t-\alpha-\beta y)]\{1 - G[(t-\alpha-\beta x) \vee (t-\alpha-\beta y)]\}\delta(x)\delta(y)dxdy \right\} + \\
&+ N^{-1}(1-\pi) \int [G(t-\alpha-\beta x) - G(t-\alpha-\beta x)^2]\delta(x)dx + o(n^-1),
\end{aligned}
$$

utilizzando gli stimatori $\hat{G}_n(t)$, $\hat{g}(t)$, $\hat{\sigma}^2$ e $\hat{\sigma_x^2}$ proposti da Wang e Dorfman (1996) si ottiene uno stimatore per la varianza proposta da Wu e Sitter (2001) per lo stimatore CD:

$$
\begin{aligned}
&\hat{V}[\hat{F}_N(t) - F_N(t)] = \\
&= n^{-1}(1-\pi)^2 \left\{ \hat{\sigma_x}^{-2}\hat{\sigma}^2 \left((N-n)^{-1} \sum_{k \in r} (x_k - \bar{x}) \hat{g}(t-\hat{\alpha}-\hat{\beta}x_k) \right)^2 + \right. \\
&\left. + (N-n)^{-2} \sum_{i \in s} \sum_{k \in r} \hat{G}_n[(t-\hat{\alpha}-\hat{\beta}x_i) \wedge (t-\hat{\alpha}-\hat{\beta}x_k)]\{1 - \hat{G}_n[(t-\hat{\alpha}-\hat{\beta}x_i) \vee (t-\hat{\alpha}-\hat{\beta}x_k)]\} \right\} + \\
&+ N^{-1}(1-\pi)(N-n)^{-1} \sum_{k \in r} [\hat{G}_n(t-\hat{\alpha}-\hat{\beta}x_k) - \hat{G}_n(t-\hat{\alpha}-\hat{\beta}x_k)^2].
\end{aligned}
$$

Lo stimatore proposto presenta diverse criticità[8]. Innanzitutto lo stimatore dipende dalla stima kernel della funzione di densità g dell'errore del modello di regressione (il modello di superpopolazione scelto), che a sua volta dipende dal parametro di bandwidth h e dalla funzione kernel $K(\cdot)$, per i quali non esiste un unico criterio ottimale di scelta. Il fatto di dover fare una scelta arbitraria per alcuni parametri nel processo di stima della varianza costituisce un problema. Un altro inconveniente dello stimatore proposto da Wang e Dorfman (1996), e anche della variante proposta da Wu e Sitter (2001), è quello di dipendere dal modello lineare (3.14). Lo stimatore della varianza (Wang e Dorfman, 1996; Wu e Sitter, 2001) si basa sulla varianza della stima CD per la funzione di ripartizione derivata da Chambers *e altri* (1992) che è valida solo nel caso in cui il modello di superpopolazione per la variabile di studio sia quello lineare semplice espresso dalla (3.14). Se si ipotizza un altro modello di superpopolazione allora la varianza analitica proposta da Chambers *e altri* (1992) non è più valida e deve essere determinata una nuova formulazione analitica della varianza dello stimatore CD. Per risolvere questo problema sono stati proposti due stimatori alternativi. Il primo si basa sul metodo jackknife ed è stato proposto da Wu e Sitter (2001), il secondo si basa sul metodo bootstrap ed è stato proposto da Lombardia *e altri* (2003). Lo stimatore proposto da Lombardia *e altri* (2003) sarà trattato

[8] Si considera la proposta di Wu e Sitter (2001) come una variante dello stimatore proposto da Wang e Dorfman (1996).

in modo dettagliato nel capitolo 4, mentre lo stimatore proposto da Wu e Sitter (2001) sarà trattato sinteticamente in questa sezione. Il motivo di tale discriminazione dipende dal fatto che, nell'ambito della stima per piccole aree basata sul modello di regressione M-quantile e solo per quanto riguarda i casi trattati in questo testo,con il metodo di stima della varianza dello stimatore CD della funzione di ripartizione proposto da Wu e Sitter (2001) non si sono ottenuti buoni risultati empirici.

La procedura jackknife per la stima della varianza dello stimatore CD proposta da Wu e Sitter (2001) si basa sulla scomposizione in due parti della varianza stessa. Si consideri la differenza tra lo stimatore CD, $\hat{F}_N(t)$, e la vera funzione di ripartizione, $F_N(t)$, considerando il modello (3.14) come modello di superpopolazione:

$$\hat{F}_N(t) - F_N(t) = (nN)^{-1} \sum_{k \in r} \sum_{i \in s} I(y_i + \hat{\beta}(x_k - x_i) \leq t) - N^{-1} \sum_{k \in r} I(y_i \leq t).$$

Posto

$$\begin{aligned} V_1 &= V\left[n^{-1} \sum_{i \in s} N^{-1} \sum_{k \in r} I(y_i + \hat{\beta}(x_k - x_i) \leq t) \right] \\ V_2 &= V\left[N^{-1} \sum_{k \in r} I(y_i \leq t) \right] = N^{-2} \sum_{k \in r} [G(t - \alpha - \beta x_k) - G(t - \alpha - \beta x_k)^2], \end{aligned}$$

la varianza dello stimatore CD risulta essere $V[\hat{F}_N(t) - F_N(t)] = V_1 + V_2$. Si consideri le condizioni **C-i**, **C-ii** e

C-iv. $N^{-1} \sum_{i \in \Omega} x_i^2$ esiste finito.

C-v. $E[X^{1/q}]$ esiste finito per $q \in (0, 1/4)$, dove X indica la variabile ausiliaria[9].

C-vi. Per un t fissato esistono finite le grandezze $N^{-1} \sum_{i \in \Omega} g(t - \alpha - \beta x_i)$ e $N^{-1} \sum_{i \in \Omega} x_i g(t - \alpha - \beta x_i)$.

Utilizzando le condizioni elencate Wu e Sitter (2001) dimostrano il seguente teorema:

Teorema 3.3.1 *Sia*

$$\begin{aligned} \hat{V}_1 &= n^{-1}(n-1) \sum_{i \in s} (F_i^* - \bar{F}^*)^2 \\ \hat{V}_2 &= N^{-2} \sum_{k \in r} [\hat{G}_n(t - \hat{\alpha} - \hat{\beta} x_k) - \hat{G}_n(t - \hat{\alpha} - \hat{\beta} x_k)^2], \end{aligned}$$

dove

$$F_i^* = (n-1)^{-1} \sum_{i \in s_l} \left[N^{-1} \sum_{k \in r} I(y_l + \hat{\beta}_l(x_k - x_i)) \right] \quad e$$

[9] Questa condizione è più stringente rispetto alla condizione posta in Wu e Sitter (2001), tuttavia risulta più leggibile e non perde di generalità.

$$\bar{F}^* = n^{-1} \sum_{i \in s} F_i^*,$$

s_l è il campione s esclusa l'unità l-esima e $\hat{\beta}_l$ è il coefficiente di regressione calcolato sul campione s_l. Se valgono le condizioni ***C-i,C-ii,C-iv,C-v,C-vi*** *e il modello di superpopolazione* (3.14)*, allora*

$$\widehat{Var}[\hat{F}_N(t) - F_N(t)] = \hat{V}_1 + \hat{V}_2$$

è uno stimatore consistente di $Var[\hat{F}_N(t) - F_N(t)]$ *e* $\widehat{Var}[\hat{F}_N(t) - F_N(t)]^{-1/2}[\hat{F}_N(t) - F_N(t)] \xrightarrow{d.} N(0,1)$.

L'approccio jackknife alla stima della varianza dello stimatore CD è stato presentato utilizzando il modello di superpopolazione (3.14). Ciò nonostante una sua estensione ad un qualsiasi modello di superpopolazione risulta di facile implementazione superando di fatto il problema della stima della varianza analitica, completamente legata al modello (3.14).

Lo Stimatore Chambers-Dunstan Applicato alla Stima per Piccole Aree

Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita e $\boldsymbol{y} = [y_1, \ldots, y_N]'$ il vettore della variabile d'interesse (per tutte le unità della popolazione Ω). Si consideri un campione casuale $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^m n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^m N_i - n_i = N - n$.

Si ipotizzi che le variabili ausiliarie $\{X_1, \ldots, X_p\}$ siano note per tutte le unità della popolazione Ω in modo certo e senza errore, e sia $\boldsymbol{x}_{ji} = [x_{j1}, \ldots, x_{jp}]'$ il vettore delle variabili ausiliarie relativo all'unità j-esima appartenente all'area i. In queste condizioni si possono applicare quei metodi di stima per piccole aree denominati metodi di stima *a livello di unità (o unit level, in inglese)*.

Il modello prevalentemente usato nella stima per piccole aree a livello di unità è il modello lineare ad effetti misti a cui si affianca il nuovo approccio basato sul modello di regressione M-quantile[10]. Entrambi sono stati presentati nel capitolo 2. Riepilogando, per predire i valori della variabile di studio, Y, non osservati, $\boldsymbol{y}_r$, si propongono i seguenti modelli:

$$\begin{aligned} \hat{y}_{ji}^{mm} &= \boldsymbol{x}_{ji}' \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i \\ \hat{y}_{ji}^{mq} &= \boldsymbol{x}_{ji}' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i), \end{aligned} \tag{3.23}$$

dove $\hat{y}_{ji}^{mm}$ è il valore predetto per la variabile Y per l'unità (anche non osservata) j-esima appartenente all'area i, ottenuto utilizzando il modello lineare ad effetti misti (si veda in proposito il capitolo 2)[11], mentre $\hat{y}_{ji}^{mq}$ rappresenta la stessa grandezza, ottenuta utilizzando il modello di regressione M-quantile (capitolo 2)[12].

[10] Il modello lineare ad effetti misti è prevalentemente usato in tutti gli approcci nell'ambito della stima per piccole aree. Tuttavia nel testo si specifica l'utilizzo del modello di stima per piccole aree a livello di unità poiché è il contesto che si considera nella stesura della tesi.

[11] Per una trattazione esauriente sull'argomento si consulti Rao (2003).

[12] Per approfondimenti si consulti Breckling e Chambers (1988) e Chambers e Tzavidis (2006).

Il modo più intuitivo, denominato *naïve*, per stimare la funzione di riaprtizione è quello di utilizzare li valori predetti $\hat{y}_{ji}$ al posto dei valori incogniti $y_{ji}, j \in r_i$ nella formula (3.16), come suggerito nella (3.17). Si ipotizzi che esista un modello di superpopolazione per la variabile di studio Y della popolazione finita Ω. Si ipotizzi che valga come modello di superpopolazione il modello lineare ad effetti misti o il modello di regressione M-quantile, ovvero:

$$\begin{aligned} y_{ji} &= \boldsymbol{x}'_{ji}\boldsymbol{\beta} + u_i + \varepsilon_{ji} \text{ oppure} \\ y_{ji} &= \boldsymbol{x}'_{ji}\boldsymbol{\beta}_\psi(\theta_i) + \varepsilon_{ji}, \end{aligned}$$

dove y_{ji} è l'unità j-esima appartenente all'area i, $\boldsymbol{x}_{ji}$ è il vettore delle variabili ausiliarie per l'unità j-esima dell'area i (che ricordiamo noto senza errore per tutte le unità della popolazione), $\boldsymbol{\beta}$ e u_i sono rispettivamente il vettore dei coefficienti di regressione degli effetti fissi e l'effetto casuale dell'area i-esima nel modello lineare ad effetti misti. θ_i è l'M-quantile medio per l'area i-esima, $\boldsymbol{\beta}_\psi$ è il vettore dei coefficienti di regressione nella regressione M-quantile e, in entrambi i casi, ε_{ji} è un errore non osservabile e incognito riferito all'unità j-esima dell'area i, distribuito normalmente con media 0 e varianza σ^2. Nel caso del modello di regressione M-quantile non è necessario assumere che ε_{ji} sia distribuito normalmente. Il valore predetto per la variabile Y per le unità non campionate ($j \in r_i$) nelle m aree per questi due modelli è quello specificato nella (3.23). Considerando queste premesse, lo stimatore *naïve* della funzione di ripartizione per la piccola area i è:

$$\tilde{F}_{i,N}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_i \le t) + \sum_{k\in r_i} I(\hat{y}_k \le t)\right], \tag{3.24}$$

dove $\hat{y_k}$ è uno dei due predittori della variabile di studio Y espressi nella (3.23) per l'unità non campionata k-esima appartenente all'area i. Lo stimatore *naïve* della funzione di ripartizione per piccola area ($\tilde{F}_{i,N}(t)$) è uno stimatore non corretto, esattamente come nel caso dello stimatore *naïve* della funzione di ripartizione per la popolazione ($\tilde{F}_N(t)$), presentato nella (3.17). Un approccio per ottenere una stima corretta della funzione di ripartizione per piccola area è quello di usare lo stimatore proposto da Chambers e Dunstan (1986) opportunamente adattato alla stima per piccole aree. Una prima proposta per la stima della funzione di ripartizione per piccola area è stata fatta in Tzavidis *e altri* (2008a). Nel caso in cui si utilizzi il modello lineare ad effetti misti come predittore per la Y lo stimatore CD per la piccola area j si ottiene basandosi sulla formulazione generale dello stimatore CD (3.18). Si ponga $T(y, \lambda = \boldsymbol{\beta}, u_i) = W_{ji} = y_{ji} - \boldsymbol{x}'_{ji}\boldsymbol{\beta} + u_i$. La stima dei parametri $\boldsymbol{\beta}$ e u_i si ottiene seguendo quanto detto nel capitolo 2. Sia ϵ_{ji}^{mm} il residuo del modello lineare ad effetti misti:

$$\epsilon_{ji}^{mm} = y_{ji} - \boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}} - \hat{u}_i, \quad j \in s_i, \quad i = 1, \ldots, m.$$

La funzione di ripartizione G della variabile casuale W è stimata con la funzione di ripartizione empirica dei residui dell'area j-esima, ovvero $\hat{G}_{n,i}(t) = n_i^{-1}\sum_{j\in s_i} I(\epsilon_{ji}^{mm} \le t)$, mentre $T(y, \lambda = \boldsymbol{\beta}, u_i)$ è stimato da $T(y, \lambda_n = \hat{\boldsymbol{\beta}}, \hat{u}_i) = \epsilon_{ji}^{mm}$. Sostituendo nella (3.18) $T(y, \lambda)$ con $T(y, \lambda_n)$, e $G_i(t)$ con $\hat{G}_{n,i}(t)$ si ottiene lo stimatore CD della funzione di ripartizione per piccola area basato sul modello lineare ad effetti misti:

$$\hat{F}_{N,i}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ji} \le t) + \sum_{k\in r_i} \hat{G}_{n,i}(t - \boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}} - \hat{u}_i)\right]. \tag{3.25}$$

Procedendo analogamente si ottiene lo stimatore CD basato sul modello di regressione M-quantile. Si ponga $T(y, \lambda = \boldsymbol{\beta}_\psi, \theta_i) = W_{ji} = y_{ji} - \boldsymbol{x}'_{ji}\boldsymbol{\beta}_\psi(\theta_i)$ e sia

$$\epsilon^{mq}_{ji} = y_{ji} - \boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i), \quad j \in s_i, \quad i = 1, \ldots, m,$$

il residuo del modello di regressione M-quantile. Ottenuta una stima dei parametri $\boldsymbol{\beta}_\psi$ e θ_i con le tecniche presentate nel capitolo 2, si ottiene una stima di $T(y, \lambda = \boldsymbol{\beta}, \theta_i)$ con $T(y, \lambda_n = \hat{\boldsymbol{\beta}}, \hat{\theta}_i) = \epsilon^{mq}_{ji}$. La funzione di ripartizione $G(t)$ della variabile casuale W è stimata con la funzione di ripartizione empirica del residuo del modello di regressione M-quantile dell'area i, $\hat{G}_{n,i}\epsilon^{mq}_{ji}$. Sostituendo nella (3.18) $T(y, \lambda)$ con $T(y, \lambda_n)$, e $G_i(t)$ con $\hat{G}_{n,i}(t)$ si ottiene lo stimatore CD della funzione di ripartizione per piccola area basato sul modello sul modello di regressione M-quantile:

$$\hat{F}_{N,i}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ji} \leq t) + \sum_{k\in r_i} \hat{G}_{n,i}(t - \boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i))\right]. \tag{3.26}$$

Per un'interpretazione più agevole dello stimatore CD per la funzione di ripartizione si può riscrivere la (3.26) come segue[13]:

$$\hat{F}_{N,i}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ji} \leq t) + \sum_{k\in r_i} n_i^{-1} \sum_{j\in s_i} I(\boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + \epsilon_{ji} \leq t)\right], \tag{3.27}$$

dove $\sum_{k\in r_i} n_i^{-1} \sum_{j\in s_i} I(\boldsymbol{x}'_{ji}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + \epsilon_{ji} \leq t)$ è un predittore rispetto al modello di regressione M-quantile della quantità incognita $\sum_{k\in r_i} I(y_{ki} \leq t)$.

Si è visto come si ottiene lo stimatore CD della funzione di ripartizione per piccola area partendo da un modello predittivo per la variabile di interesse adeguato al caso delle piccole aree. Ottenuta una stima dei parametri, λ, del modello, si ricorre alla trasformazione $T(y, \lambda_n)$ ed alla stima della sua funzione di ripartizione ($\hat{G}_n$(t)) suggerite da Chambers e Dunstan (1986) per ottenere uno stimatore, asintoticamente corretto rispetto al modello, della funzione di ripartizione per popolazioni finite.

Per la stima dell'errore quadratico medio della funzione di ripartizione non esiste una formulazione generale facilmente adattabile ad un qualsiasi modello. Infatti, la formulazione per la stima dell'errore quadratico medio proposta nella (3.22) vale solo se il modello di superpopolazione è il modello di regressione lineare. Tale modello (il (3.14)) non si adatta bene al problema della stima per piccole aree poiché non è possibile catturare la variabilità tra le aree. Dunque non è utile un'estensione della (3.22) nel caso della stima per piccole aree. Al momento della stesura di questa tesi non ci sono in letteratura proposte per la stima dell'errore quadratico medio dello stimatore CD per piccola area. Si rimanda il lettore al capitolo 4 per una proposta di uno stimatore per l'errore quadratico medio dello stimatore CD per piccola area.

3.3.3 Lo Stimatore Rao-Kovar-Mantel

Rao *e altri* (1990) propongono uno stimatore per la funzione di ripartizione non distorto sia rispetto al modello, sia rispetto al disegno.

Si consideri la popolazione finita Ω formata da N unità, $i = (1, \ldots, N)$. Si consideri il carattere continuo Y della popolazione Ω, $\{y_1, \ldots, y_N\}$. Si estragga un campione s da Ω di $n \leq N$ unità e

[13] E' evidente che una scrittura del genere è possibile anche per la (3.25). Tuttavia il focus di questo lavoro resta la stima della funzione di ripartizione con il metodo CD utilizzando il modello di regressione M-quantile.

su ogni unità si misuri il carattere Y. Condizionandoci al campione estratto, il carattere Y è noto per le n unità appartenenti al campione s, ovvero $\{y_1, \ldots, y_n\}$ è noto. Sia π_i la probabilità di includere l'unità i-esima della popolazione Ω nel campione s. Uno stimatore non distorto rispetto al disegno è:

$$\hat{F}_N(t) = \sum_{i \in s} \pi_i^{-1} I(y_i \leq t) \left(\sum_{i \in s} \pi_i^{-1} \right)^{-1} . \tag{3.28}$$

Stimatori alternativi alla (3.28) sono stati proposti da Kuk (1988). Tuttavia la formulazione per la stima della funzione di ripartizione (3.28) non include eventuali variabili ausiliarie. In molti problemi applicati alcune variabili sono note per tutte le unità della popolazione (reperibili ad esempio fonti censuarie e archivi amministrativi esaustivi per la popolazione di riferimento) ed un loro utilizzo nella procedura di stima è auspicabile. Nell'ambito della stima per piccole aree l'utilizzo delle variabili ausiliarie è determinante. Infatti, sfruttando una relazione esistente, o ipotizzata, tra la variabile di interesse e le variabili ausiliarie si cerca di ottenere degli stimatori con una variabilità contenuta nonostante la ridotta numerosità campionaria nelle piccole aree.

Uno stimatore basato su disegno che utilizza variabili ausiliarie è lo stimatore rapporto della funzione di ripartizione. Sia $\hat{R} = (\sum_{i \in s} y_i/\pi_i)/(\sum_{i \in s} x_i/\pi_i)^{-1}$ lo stimatore rapporto della variabile casuale $R = Y/X$, dove Y indica la variabile di interesse e X la variabile ausiliaria. $\hat{R}$ è uno stimatore consistente basato su disegno. Lo stimatore rapporto della funzione di ripartizione è:

$$\hat{F}_N^r(t) = N^{-1} \left[\sum_{i \in s} \pi_i^{-1} I(y_i \leq t) \right] \left[\sum_{i \in s} \pi_i^{-1} I(\hat{R} x_i \leq t) \right]^{-1} \left[\sum_{i \in \Omega} I(\hat{R} x_i \leq t) \right] . \tag{3.29}$$

Nel caso in cui la variabile di interesse Y sia approssimativamente proporzionale alla variabile ausiliaria X si dimostra (Rao *e altri*, 1990) che lo stimatore $\hat{F}_N^r(t)$ (3.29) è più efficiente dello stimatore proposto nella (3.28).

In alternativa allo stimatore rapporto si può utilizzare lo stimatore differenza della funzione di ripartizione:

$$\hat{F}_N^d(t) = N^{-1} \left[\sum_{i \in s} \pi_i^{-1} I(y_i \leq t) + \sum_{i \in \Omega} I(\hat{R} x_i \leq t) - \sum_{i \in s} \pi_i^{-1} I(\hat{R} x_i \leq t) \right] ,$$

che ha il vantaggio di non risentire della distorsione dello stimatore rapporto, soprattutto nel caso di campioni di piccole dimensioni. $\hat{F}_N^d(t)$ è uno stimatore non distorto rispetto al disegno (Rao *e altri*, 1990).

Se si ipotizza che la variabile di interesse Y è generata da un modello di superpopolazione, gli stimatori $\hat{F}_N^r(t)$ e $\hat{F}_N^d(t)$ risultano distorti rispetto al modello.

Si consideri il modello di regressione lineare semplice, $y_i = \alpha + \beta x_i + \varepsilon_i$, come modello di superpopolazione. Si consideri la popolazione finita Ω formata da N unità, $i = 1, \ldots, N$. Si consideri i caratteri Y e X della popolazione Ω, con Y carattere continuo. Per le N unità della popolazione si ha la coppia (y_i, x_i), $i = 1, \ldots, N$. Si ipotizzi di conoscere in modo certo e senza errore il carattere X per tutte le unità della popolazione e di non conoscere il carattere Y. Si estragga un campione s da Ω di $n \leq N$ unità e su ogni unità si misuri il carattere Y. Condizionandoci al campione estratto, il carattere Y è noto per le n unità appartenenti al campione s, ovvero $\{y_1, \ldots, y_n\}$ è noto. Sia π_i la probabilità di includere l'unità i-esima della popolazione Ω nel campione s. Lo stimatore differenza della funzione di ripartizione proposto da Rao *e altri* (1990) è

$$F_N^{RKM}(t) = N^{-1}\left[\sum_{i\in s}\pi_i^{-1}I(y_i \le t) + \sum_{i\in\Omega}G_i - \sum_{i\in s}\pi_i^{-1}G_i\right]. \quad (3.30)$$

Lo stimatore $F_N^{RKM}(t)$ è sia non distorto rispetto al modello sia non distorto rispetto al disegno (Rao *e altri*, 1990). La quantità G_i, secondo quanto ipotizzato, è:

$$G_i = N^{-1}\sum_{k\in\Omega} I(Rx_i + \gamma_k \le t),$$

dove R è il (vero) rapporto tra la variabile di studio Y, generata dal modello di superpopolazione $y_i = \alpha + \beta x_i + \varepsilon_i$, e la variabile ausiliaria X, e $\gamma_k = y_k - Rx_k$.

Lo stimatore (3.30) non può essere calcolato poiché G_i richiede di conoscere il vero valore di R e il valore del carattere di interesse Y per tutte le unità della popolazione. Uno stimatore per la (3.30), d'ora in avanti stimatore RKM, proposto da Rao *e altri* (1990) è il seguente:

$$\hat{F}_N^{RKM}(t) = N^{-1}\left[\sum_{i\in s}\pi_i^{-1}I(y_i \le t) + \sum_{i\in\Omega}\hat{G}_i - \sum_{i\in s}\pi_i^{-1}\hat{G}_{ic}\right], \quad (3.31)$$

dove $\hat{G}_i$ e $\hat{G}_{ic}$, con $i \in s$, sono stimatori di G_i asintoticamente corretti rispetto al disegno (Rao *e altri*, 1990):

$$\begin{aligned}
\hat{G}_i &= \left(\sum_{i\in s}\pi_i^{-1}\right)^{-1}\left[\sum_{j\in s}\pi_j^{-1}I(\hat{R}x_i + \hat{\gamma}_j \le t)\right] \\
\hat{G}_{ic} &= \left(\sum_{j\in s}\pi_i/\pi_{ji}\right)^{-1}\left[\sum_{j\in s}\pi_i/\pi_{ji}I(\hat{R}x_i + \hat{\gamma}_j \le t)\right],
\end{aligned}$$

dove $\hat{\gamma}_j = y_j - \hat{R}x_j$ con $j \in s$, π_{ji} è la probabilità di includere nel campione s sia l'unità i, sia l'unità j e π_{ji}/π_i è la probabilità condizionata di inclusione nel campione s delle unità i e j dato che l'unità i appartiene al campione s. Lo stimatore $\hat{F}_N^{RKM}(t)$ (3.31) è asintoticamente non distorto per la popolazione finita Ω, sia rispetto al modello sia rispetto al disegno (Rao *e altri*, 1990).

La varianza dello stimatore RKM ($\hat{F}_N^{RKM}(t)$) proposta da Rao *e altri* (1990) è la seguente:

$$V(\hat{F}_{RKM}(t)) = N^{-2}V(I(y_i \le t) - G_i),$$

e un suo stimatore è:

$$\hat{V}(\hat{F}_{RKM}(t)) = \sum_{i<j\in s}(\pi_i\pi_j - \pi_{ji})\pi_{ji}^{-1}[I(y_i \le t) - \hat{G}_{ji}(j)\pi_i^{-1} - I(y_i \le t) - \hat{G}_{jc}(i)\pi_j^{-1}]^2, \quad (3.32)$$

dove

$$\hat{G}_{ic}(j) = \left(\sum_{k\in s}\pi_{ji}/\pi_{jki}\right)^{-1}\left[\sum_{k\in s}(\pi_{ji}/\pi_{jki})I(\hat{R}x_i - \hat{\gamma}_k \le t)\right],$$

dove π_{jki} è la probabilità di includere nel campione s le unità i, j e k, mentre $\pi_{jki}/\pi ji$ è la probabilità condizionata di inclusione nel campione s delle unità i, j e k dato che le unità i e j appartengono al campione s. Lo stimatore $\hat{G}_{ic}(j)$ condizionato a $(i, j \in s)$ è asintoticamente corretto per G_i rispetto al disegno (Rao *e altri*, 1990).

Secondo le simulazioni fatte da Rao *e altri* (1990) lo stimatore RKM ha ottenuto delle performance migliori rispetto allo stimatore CD, almeno per quanto concerne l'errore relativo medio (3.10). Tuttavia, sempre secondo i risultati ottenuti da Rao *e altri* (1990), lo stimatore CD è risultato più preciso nel caso di piccoli campioni.

La stima della varianza per lo stimatore RKM è più facile da ottenere rispetto alla varianza dello stimatore CD, tuttavia lo stimatore RKM è limitato all'utilizzo di una sola variabile ausiliaria. Inoltre, è necessario conoscere la probabilità di inclusione del secondo ordine per poter utilizzare lo stimatore RKM mentre è necessaria quella del terzo ordine per la stima della varianza dello stimatore RKM. Conoscere le probabilità di inclusione del secondo e terzo ordine non è scontato, fatta eccezione per il campione casuale semplice[14]. Rao *e altri* (1990) per sopperire a questo inconveniente rimandano ad una proposta di Chao (1982) per calcolare la probabilità di inclusione fino al terzo ordine per campioni probabilistici[15].

Lo Stimatore Rao-Kovar-Mantel Applicato alla Stima per Piccole Aree

Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita e sia $\boldsymbol{y} = [y_1, \ldots, y_N]'$ il vettore della variabile d'interesse per tutte le unità della popolazione Ω. Si consideri un campione $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^{m} n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^{m} N_i - n_i = N - n$.

Si ipotizzi che le variabili ausiliarie $\{X_1, \ldots, X_p\}$ siano note per tutte le unità della popolazione Ω in modo certo e senza errore, e sia $\boldsymbol{x}_{ji} = [x_{j1}, \ldots, x_{jp}]'$ il vettore delle variabili ausiliarie relativo all'unità j-esima appartenente all'area i. Si indichi con $\hat{y}_{ji}^{mm}$ e $\hat{y}_{ji}^{mq}$ (3.23) il predittore per la variabile di interesse ottenuto rispettivamente con il modello lineare ad effetti misti o il modello di regressione M-quantile. Si ipotizzi che s sia un campione casuale semplice estratto dalla popolazione finita Ω. In questo caso lo stimatore RKM per la piccola area i è:

$$\begin{aligned}\hat{F}_{N,i}^{RKM}(t) = n_i^{-1} \sum_{j \in s_i} I(y_{ji} \leq t) + N_i^{-1} n_i^{-1} \sum_{k \in r_i} \sum_{j \in s_i} I(\hat{y}_{ki} + (y_{ji} - \hat{y}_{ji}) \leq t) - \\ -(n_i^{-1} - N_i^{-1}) n_i^{-1} \sum_{k \in r_i} \sum_{j \in s_i} I(\hat{y}_{ki} + (y_{ji} - \hat{y}_{ji}) \leq t),\end{aligned} \tag{3.33}$$

dove $\hat{y}_{ki}$ è il predittore $\hat{y}_{ji}^{mm}$ oppure $\hat{y}_{ji}^{mq}$ (3.23). Questa versione dello stimatore RKM adattata al problema della stima per piccole aree è stata proposta in Tzavidis *e altri* (2008b). Uno stimatore della variabilità per la (3.33) può essere facilmente ottenuto dalla (3.32). La stima della variabilità dello stimatore RKM proposto (3.33) non è trattata in questo lavoro[16].

[14] E, in genere, per qui disegni campionari che associano ad ogni unità della popolazione uguale probabilità di appartenenza al campione.

[15] Si fa riferimento, in particolare, ai disegni di campionamento a probabilità variabile.

[16] Non sono noti, alla data di pubblicazione di questo testo, lavori inerenti la stima della variabilità dello stimatore RKM nell'ambito della statistica per piccole aree.

E' stato fatto uno studio di simulazione da Chambers *e altri* (1992) per confrontare lo stimatore CD con lo stimatore RKM nell'ambito della stima per piccole aree. I risultati che Chambers *e altri* (1992) hanno ottenuto mostrano che in termini di variabilità dello stimatore non si riscontra una maggior efficienza dello stimatore RKM verso lo stimatore CD. Nel caso in cui il modello di superpopolazione per la variabile d'interesse sia mal specificato lo stimatore CD è meno efficiente dello stimatore RKM e viceversa. Questo risultato, relativo agli stimatori (3.33) e (3.25) (o (3.26)), relativi alla stima per piccole aree, riportato da Chambers *e altri* (1992) è in linea con i risultati ottenuti da Rao *e altri* (1990).

3.4 La Stima dei Quantili Tramite la Stima della Funzione di Ripartizione

In questo paragrafo si descrive un modo per ottenere la stima dei quantili di una distribuzione partendo dalla stima della funzione di ripartizione, comunque sia stata ottenuta.

Sfruttando la notazione integrale di Riemann-Stieltjes, la definizione di un quantile per la distribuzione di una generica variabile casuale X di ordine τ, ($\tau \in [0,1]$), $q_X(\tau)$ è:

$$\int_{-\infty}^{q_X(\tau)} dF(x) = \tau, \tag{3.34}$$

dove $F(x)$ è la funzione di ripartizione e $dF(x)/dx = f(x)$, con $f(x)$ funzione di densità della variabile casuale X. Sostituendo nella (3.34) $F(x)$ con una sua stima $\hat{F}(x)$ si ottiene la stima di un quantile di ordine τ, $\hat{q}_X(\tau)$:

$$\int_{-\infty}^{\hat{q}_X(\tau)} d\hat{F}(x) = \tau. \tag{3.35}$$

L'integrale nella (3.35) si può risolvere numericamente.

Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita e sia $\boldsymbol{y} = [y_1, \ldots, y_N]'$ il vettore della variabile d'interesse per tutte le unità della popolazione Ω. Si consideri un campione $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^{m} n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^{m} N_i - n_i = N - n$.

Si ipotizzi che le variabili ausiliarie $\{X_1, \ldots, X_p\}$ siano note per tutte le unità della popolazione Ω in modo certo e senza errore, e sia $\boldsymbol{x}_{ji} = [x_{j1}, \ldots, x_{jp}]'$ il vettore delle variabili ausiliarie relativo all'unità j-esima appartenente all'area i. Si indichi con $\hat{y}_{ji}^{mm}$ e $\hat{y}_{ji}^{mq}$ (3.23) il predittore per la variabile di interesse ottenuto rispettivamente con il modello lineare ad effetti misti o il modello di regressione M-quantile. Sia $\hat{F}_{N,i}^{CD}$ lo stimatore CD per la funzione di ripartizione della variabile di interesse Y nella piccola are i, ottenuto usando il predittore $\hat{y}_{ji}^{mm}$ o $\hat{y}_{ji}^{mq}$ (3.23). Lo stimatore del quantile di ordine τ per la variabile di interesse Y nella piccola area i, $\hat{q}_{i,Y}(\tau)$ è dato da:

$$\int_{-\infty}^{\hat{q}_{i,Y}(\tau)} d\hat{F}_{N,i}^{CD}(y) = \tau. \tag{3.36}$$

La stima del quantile $q_{i,Y}(\tau)$ si può ottenere alternativamente utilizzando lo stimatore RKM per piccole aree (3.33). L'integrale nella (3.36) si risolve numericamente. Per le stime dei quantili presentate

in questo testo, ottenute con la (3.36), utilizzando come predittore per la variabile di interesse per l'unità j-esima appartenente all'area i il predittore $\hat{y}_{ji}^{mq}$ (3.23), è stato utilizzato un algoritmo scritto ad hoc nell'ambiente statistico *R*[17].

3.5 Un Confronto tra lo Stimatore CD e lo Stimatore Naïve e Campionario per la Stima dei Quantili nell'Ambito della Stima per Piccole Aree

In questo paragrafo si presenta un confronto tra diversi metodi di stima dei quantili fatto tramite simulazione Monte Carlo. Viste le simulazioni fatte da Chambers *e altri* (1992) e Rao *e altri* (1990), basate soprattutto sull'efficienza degli stimatori, nelle simulazioni qui presentate si è posta l'attenzione sulla correttezza degli stimatori. Sono state fatte due simulazioni, una basata su modello e una basata su disegno.

Nella simulazioni basata su modello è stata generata una popolazione finita Ω per ogni iterazione Monte Carlo. La popolazione finita Ω di N unità è stata divisa in 30 aree di dimensione pari a $N_1, \ldots, N_{30}$ unità in modo che $\sum_{i=1}^{30} N_i = N$. Per ogni unità della popolazione Ω è stata generata la coppia di valori (y_{ji}, x_{ji}) con $i = (1, \ldots, 30)$, indice di area, e $j = (1, \ldots, N_i)$, indice di unità nell'area i. Le x_{ji} sono state generate da una distribuzione normale standard con media μ_i e varianza uguale per tutte le aree e pari a $\sigma_x^2 = 1$. Le 30 medie μ_i sono state generate da una distribuzione uniforme discreta con parametri 20, 100 con i valori μ_i tenuti fissi per tutte le iterazioni Monte Carlo. Le y_{ji} sono state generate con un modello lineare ad effetti misti: $y_{ji} = \alpha + \beta x_{ji} + u_i + e_{ji}$. I parametri α e β sono stati posti pari a 1, gli effetti casuali di area, u_i, sono stati generati da una distribuzione normale standard (con media 0 e varianza $\sigma_u^2 = 1$) e gli errori e_{ji} sono stati generati da una distribuzione normale con media 0 e varianza $\sigma_e^2 = 4$. La dimensione della popolazione nelle 30 aree è stata generata da una distribuzione uniforme discreta con parametri 50, 500 ed è stata tenuta fissa per tutte le iterazioni Monte Carlo. Dalla popolazione Ω per ogni iterazione Monte Carlo è stato estratto un campione casuale semplice s di n elementi. La dimensione campionaria, tenuta fissa nelle iterazioni Monte Carlo, è stata posta pari a 1/10 della popolazione in modo che in ogni area valgano le relazioni $n_i = 0.1N_i$ e $\sum_{i=1}^{30} n_i = n$.

Per la simulazione basata su disegno è stato usato il campione LSMS Albania del 2003. Si è trattato il campione Albania come se fosse una popolazione finita, denominata Ω_{Alb}, da cui estrarre un campione casuale semplice s_{Alb}. La popolazione Ω_{Alb} (ovvero il campione Albania LSMS del 2003), è composto da 3591 unità divise in 36 aree. L'unità di riferimento indica un nucleo familiare. In ogni area è stato estratto un campione casuale semplice di dimensione pari a 1/10 della dimensione della popolazione dell'area stessa; nelle aree in cui il numero di unità della popolazione è risultato inferiore a 50 è stato estratto un campione casuale semplice di 5 unità. Il campione finale è risultato di 398 unità. La variabile di interesse, Y, è il reddito equivalente (espresso in euro), mentre come unica variabile ausiliaria, X, è stata utilizzata la dimensione della famiglia[18]. La popolazione Ω_{Alb} è fissa nelle iterazioni Monte Carlo mentre ad ogni iterazione è stato estratto un campione s_{Alb} secondo le modalità descritte.

Siano s_i e $s_{i,Alb}$ le unità dei campioni estratti rispettivamente dalle popolazioni Ω e Ω_{Alb} nell'area i e si indichi con r_i o $r_{i,Alb}$ le $N_i - n_i$ unità non campionate nell'area i. Sia inoltre $r = N - n$

[17] http://www.r-project.org. L'algoritmo è stato gentilmente concesso dal Dott. Nikolaos Tzavidis, Center for Census and Survey Research, University of Manchester, UK.

[18] Famiglia in questo contesto è da intendersi come l'insieme degli individui che formano un nucleo. Tale concetto è generalmente espresso con il termine inglese *household*.

l'insieme delle unità non campionate della popolazione Ω e $r_{Alb} = N - n$ l'insieme delle unità non campionate della popolazione Ω_{Alb}. Per semplificare la notazione si ponga $s_{i,Alb} = s_i$, $r_{i,Alb} = r_i$, $s_{Alb} = s$ e $r_{Alb} = r$. Il pedice "Alb sarà inserito ogni qualvolta sia necessario distinguere tra le grandezze della simulazione basata su modello da quella basata su disegno.

Per entrambe le simulazioni sono state fatte 500 iterazioni Monte Carlo. Ad ogni iterazione sono stati calcolati sui campioni s, riferiti alla popolazione finita Ω, e s_{Alb}, riferiti alla popolazione finita Ω_{Alb}, le stime dei quartili (ovvero i percentili 25, 50 e 75) della variabili di interesse y_{ji} ottenute con diverse metodologie:

- lo stimatore campionario del quantile ottenuto dalla funzione di ripartizione empirica:

$$\hat{q}_{i,Y}^{camp}(\tau) \text{ tale che } \int_{-\infty}^{\hat{q}_{i,Y}^{camp}(\tau)} d\hat{F}_{n,i}^{camp}(y) = \tau,$$

  dove $\tau = \{0.25, 0.50, 0.75\}$ e

$$\hat{F}_{n,i}^{camp}(t) = N_i^{-1} \sum_{j \in s_i} I(y_{ji} \leq t)^{19};$$

- lo stimatore naïve del quantile, ottenuto dalla stima naïve della funzione di ripartizione (3.17) opportunamente adattata al caso della stima per piccole aree[20]:

$$\hat{q}_{i,Y}^{naïve}(\tau) \text{ tale che } \int_{-\infty}^{\hat{q}_{i,Y}^{naïve}(\tau)} d\hat{F}_{N,i}^{naïve}(y) = \tau,$$

  dove $\tau = \{0.25, 0.50, 0.75\}$ e

$$\hat{F}_{N,i}^{naïve}(t) = N_i^{-1} \left[\sum_{j \in s_i} I(y_{ji} \leq t) + \sum_{k \in r_i} I(\hat{y}_{ki} \leq t) \right],$$

  dove $\hat{y}_{ji}$ indica il predittore per la variabile di interesse Y ottenuto o con $\hat{y}_{ji}^{mq}$ (3.23) ed allora la stima del quantile si indica con $\hat{q}_{i,Y}^{naïve_{mq}}(\tau)$, oppure con $\hat{y}_{ji}^{mm}$ (3.23) ed allora la stima del quantile si indica con $\hat{q}_{i,Y}^{naïve_{mm}}(\tau)$;

- lo stimatore CD (Chambers-Dusntan) del quantile, ottenuto dalla stima CD della funzione di ripartizione (3.24):

$$\hat{q}_{i,Y}^{CD}(\tau) \text{ tale che } \int_{-\infty}^{\hat{q}_{i,Y}^{CD}(\tau)} d\hat{F}_{N,i}^{CD}(y) = \tau,$$

  dove $\tau = \{0.25, 0.50, 0.75\}$ e

[19] La funzione di ripartizione empirica ha funzione di stimatore per la funzione di ripartizione di una popolazione finita. $F_{n,i}^{camp} \equiv \hat{F}_{N,i}^{camp}$.

[20] Per l'adattamento dello stimatore naïve al caso delle piccole aree è sufficiente inserire opportunamente nella (3.17) uno dei predittori espressi nella (3.23).

$$\hat{F}_{N,i}^{CD}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ji} \leq t) + n_i^{-1}\sum_{k\in r_i}\sum_{j\in s_i} I(\hat{y}_{ki} + (y_{ji} - \hat{y}_{ji}) \leq t)\right],$$

dove $\hat{y}_{ji}$ indica il predittore per la variabile di interesse Y ottenuto o con $\hat{y}_{ji}^{mq}$ (3.23) ed allora la stima del quantile si indica con $\hat{q}_{i,Y}^{CD_{mq}}(\tau)$, oppure con $\hat{y}_{ji}^{mm}$ (3.23) ed allora la stima del quantile si indica con $\hat{q}_{i,Y}^{CD_{mm}}(\tau)$;

D'ora in avanti per una notazione più snella si ponga $\hat{q}_{i,Y}^{est}(\tau) = \hat{q}_i^{est}(\tau)$. Per ognuno degli stimatori presentati è stato calcolato l'errore medio relativo, RB, per ogni piccola area:

$$RB(\tau, i) = H^{-1}\sum_{h=1}^{H}\frac{\hat{q}_{i,h}^{est}(\tau) - q_{i,h}(\tau)}{q_{i,h}(\tau)},$$

e l'errore assoluto medio relativo, ARB, per ogni piccola area:

$$ARB(\tau, i) = H^{-1}\sum_{h=1}^{H}\left|\frac{\hat{q}_{i,h}^{est}(\tau) - q_{i,h}(\tau)}{q_{i,h}(\tau)}\right|,$$

dove $\tau = \{0.25, 0.50, 0.75\}$, $\hat{q}_{i,h}^{est}(\tau)$, con $est = \{camp, naïve_{mm}, naïve_{mq}, CD_{mm}, CD_{mq}\}$, è la stima del quantile di ordine τ (ottenuta con uno dei metodi precedentemente presentati) nell'area i per l'iterazione Monte Carlo h-esima, con $h = (1, \ldots, H = 500)$, mentre $q_{i,h}(\tau)$ è il valore vero del quantile di ordine τ nella popolazione di riferimento nell'iterazione h-esima della simulazione Monte Carlo. Risulta evidente che nella popolazione Ω_{Alb}, che è tenuta fissa nelle iterazioni Monte Carlo, risulta $q_i = q_{i,h}$, con $h = (1, \ldots, H = 500)$. Gli indicatori ARB e RB sono stati calcolati sia per la simulazione basata su modello sia per la simulazione basata su disegno. I risultati ottenuti dalla simulazione basata su modello sono riportati nelle tabelle 3.3 e 3.4. Per rendere più leggibili i risultati nelle tabelle 3.3 e 3.4 è stata riportata la distribuzione nelle 30 aree rispettivamente degli indici RB e ARB per gli stimatori proposti.

Tabella 3.3: Distribuzione nelle 30 aree della popolazione Ω dell'indice RB (valori %)

τ	RB	$camp$	$naïve_{mm}$	$naïve_{mq}$	CD_{mm}	CD_{mq}
	Min.	-0.01	0.77	0.77	-0.10	-0.10
	Qu. 1	0.08	1.00	1.00	-0.01	-0.01
τ	Mediana	0.14	1.38	1.38	0.02	0.02
0.25	Media	0.18	1.63	1.64	0.03	0.03
	Qu. 3	0.23	1.85	1.84	0.05	0.06
	Max.	0.57	3.77	3.68	0.30	0.28
	Min.	-0.17	-0.12	-0.11	-0.17	-0.15
	Qu. 1	-0.06	-0.02	-0.02	-0.03	-0.02
τ	Mediana	-0.01	0.00	0.00	-0.01	-0.01
0.50	Media	-0.01	0.00	0.00	0.00	0.00
	Qu. 3	0.03	0.02	0.03	0.02	0.03
	Max.	0.17	0.08	0.09	0.13	0.11
	Min.	-0.58	-3.26	-3.34	-0.19	-0.19
	Qu. 1	-0.26	-1.65	-1.68	-0.07	-0.08
τ	Mediana	-0.13	-1.28	-1.30	-0.01	-0.01
0.75	Media	-0.19	-1.52	-1.52	-0.04	-0.04
	Qu. 3	-0.07	-0.95	-0.94	0.00	0.00
	Max.	-0.03	-0.79	-0.79	0.07	0.06

Tabella 3.4: Distribuzione nelle 30 aree della popolazione Ω dell'indice ARB (valori %)

τ	ARB	$camp$	$naïve_{mm}$	$naïve_{mq}$	CD_{mm}	CD_{mq}
	Min.	0.41	0.82	0.82	0.34	0.34
	Qu. 1	0.56	1.01	1.03	0.46	0.47
τ	Mediana	0.84	1.42	1.42	0.72	0.72
0.25	Media	1.01	1.66	1.69	0.84	0.84
	Qu. 3	1.26	1.90	1.92	1.04	1.03
	Max.	2.74	3.93	3.91	2.39	2.41
	Min.	0.37	0.26	0.37	0.30	0.30
	Qu. 1	0.51	0.36	0.47	0.42	0.42
τ	Mediana	0.76	0.55	0.62	0.64	0.64
0.50	Media	0.91	0.66	0.78	0.75	0.75
	Qu. 3	1.17	0.82	0.90	0.96	0.96
	Max.	2.38	1.91	2.02	2.10	2.10
	Min.	0.39	0.80	0.82	0.31	0.31
	Qu. 1	0.54	0.96	0.96	0.45	0.44
τ	Mediana	0.82	1.31	1.34	0.68	0.68
0.75	Media	0.95	1.55	1.57	0.79	0.79
	Qu. 3	1.19	1.70	1.75	1.00	1.00
	Max.	2.63	3.38	3.52	2.15	2.15

Gli indici ARB e RB assumono per tutti gli stimatori proposti valori molto vicini a 0, indicando una precisione altissima di questi stimatori, nel contesto presentato. Questo risultato è sicuramente atteso viste le condizioni ideali in cui si è svolta la simulazione. Soprattutto per la stima della mediana (quantile di ordine $\tau = 0.5$), si nota che sia lo stimatore naïve ($naïve_{mm}$ e $naïve_{mq}$), sia lo stimatore campionario, hanno un errore, relativo e assoluto, medio pressoché pari a 0 come avviene per lo stimatore CD (CD_{mm} e CD_{mq}). Se si considera la mediana l'errore relativo medio (RB) registrato in ogni area vale al massimo 0.76% (è il caso dello stimatore campionario), è praticamente nullo. Per il primo e terzo quartile ($\tau = 0.25$ e $\tau = 0.75$) si osserva dalla tabella 3.3 e 3.4 una precisione maggiore dello stimatore CD. Comunque l'errore relativo per gli altri stimatori è molto contenuto e nel peggiore dei casi non raggiunge il 4%. Tra i due stimatori CD, quello con predittore basato sul modello lineare ad effetti misti ($\hat{q}_i^{CD_{mm}}(\tau)$) e quello con predittore basato sul modello di regressione M-quantile ($\hat{q}_i^{CD_{mq}}(\tau)$), non ve ne è uno preponderante a livello di precisione di stima. Ciò è significativo, infatti, visto che il modello con cui sono stati generati i dati rispetta le assunzioni del modello lineare ad effetti misti, si poteva presupporre una miglior performance dello stimatore $\hat{q}_i^{CD_{mm}}(\tau)$, anche se lieve, che invece non si è verificata. Per quanto concerne l'efficienza si rimanda alla simulazione fatta da Chambers *e altri* (1992), dove in effetti, in una condizione simile a quella presentata, gli stimatori basati sul modello lineare ad effetti misti sono risultati più efficienti, anche se per grandi campioni.

I risultati ottenuti dalla simulazione basata su disegno sono riportati nelle tabelle 3.5 e 3.6. Per rendere più leggibili i risultati nelle tabelle 3.5 e 3.6 è stata riportata la distribuzione nelle 36 aree, rispettivamente, degli indici RB e ARB per gli stimatori proposti.

Tabella 3.5: Distribuzione nelle 36 aree della popolazione Ω_{Alb} dell'indice RB (valori %)

τ	RB	$camp$	$naïve_{mm}$	$naïve_{mq}$	CD_{mm}	CD_{mq}
	Min.	−0.69	2.46	11.49	−2.69	−2.55
	Qu. 1	2.74	38.46	27.20	0.89	0.82
τ	Mediana	4.63	51.53	35.48	3.38	3.60
0.25	Media	7.37	52.95	39.28	5.37	5.70
	Qu. 3	9.65	66.90	52.81	7.15	7.27
	Max.	30.13	121.80	80.10	28.54	28.77
	Min.	−7.73	−18.63	−16.61	−7.67	−7.76
	Qu. 1	−0.12	3.64	−2.88	0.35	0.11
τ	Mediana	1.20	14.68	2.29	1.66	1.21
0.50	Media	2.12	15.56	4.29	2.54	2.19
	Qu. 3	4.17	25.05	9.52	4.25	3.70
	Max.	11.78	62.66	35.81	11.55	11.56
	Min.	−13.09	−39.16	−42.36	−10.60	−10.87
	Qu. 1	−5.58	−21.93	−26.88	−3.98	−4.50
τ	Mediana	−2.65	−16.81	−23.66	−1.45	−1.80
0.75	Media	−3.16	−14.04	−22.72	−1.36	−1.72
	Qu. 3	−1.00	−6.80	−18.45	1.15	1.00
	Max.	5.53	14.80	−2.84	6.53	6.49

Tabella 3.6: Distribuzione nelle 36 aree della popolazione Ω_{Alb} dell'indice ARB (valori %)

τ	ARB	$camp$	$naïve_{mm}$	$naïve_{mq}$	CD_{mm}	CD_{mq}
	Min.	4.74	9.43	14.80	4.76	4.68
	Qu. 1	11.92	38.46	27.20	11.47	11.21
τ	Mediana	15.04	51.53	35.62	14.62	14.63
0.25	Media	17.50	53.17	39.60	16.98	16.97
	Qu. 3	23.47	66.90	52.81	21.27	21.38
	Max.	36.53	121.80	80.17	36.00	36.01
	Min.	3.98	5.27	3.36	3.93	3.99
	Qu. 1	10.86	9.52	8.34	10.76	10.37
τ	Mediana	14.18	17.37	11.30	13.92	13.99
0.50	Media	15.30	19.84	12.44	15.09	15.02
	Qu. 3	21.27	25.11	14.30	20.62	20.78
	Max.	26.21	62.66	35.85	25.13	25.29
	Min.	5.83	6.95	8.34	5.99	5.93
	Qu. 1	11.90	10.78	18.80	11.93	12.10
τ	Mediana	14.58	16.84	23.70	15.17	15.16
0.75	Media	15.64	17.90	23.39	15.49	15.64
	Qu. 3	18.79	21.94	27.03	17.81	18.07
	Max.	33.20	39.16	42.40	32.31	32.66

Dalle tabelle 3.5 e 3.6 si vede che lo stimatore CD è migliore, in termini di precisione, rispetto agli altri stimatori. Lo stimatore campionario ha risultati di poco peggiori rispetto allo stimatore CD, mentre lo stimatore naïve è nettamente peggiore. Tra le due alternative di stimatori CD, quello con predittore basato sul modello lineare ad effetti misti ($\hat{q}_i^{CD_{mm}}(\tau)$) e quello con predittore basato sul modello di regressione M-quantile ($\hat{q}_i^{CD_{mq}}(\tau)$), non si nota nessuna differenza rilevante. Tra i due stimatori naïve, quello con predittore basato sul modello lineare ad effetti misti ($\hat{q}_i^{naïve_{mm}}(\tau)$) e quello con predittore basato sul modello di regressione M-quantile ($\hat{q}_i^{naïve_{mq}}(\tau)$), invece, si nota una performance migliore dello stimatore $\hat{q}_i^{naïve_{mq}}(\tau)$ per la mediana e il primo quartile, mentre per il terzo quartile lo stimatore $\hat{q}_i^{naïve_{mm}}(\tau)$ risulta essere più preciso dello stimatore $\hat{q}_i^{naïve_{mq}}(\tau)$. La media dell'indice RB degli stimatori $\hat{q}_i^{CD_{mq}}(\tau)$ e $\hat{q}_i^{CD_{mm}}(\tau)$ tra le aree è al massimo di poco inferiore al 6% (è il caso del primo quartile), mentre la media dell'indice ARB degli stimatori $\hat{q}_i^{CD_{mq}}(\tau)$ e $\hat{q}_i^{CD_{mm}}(\tau)$ tra le aree è al massimo di poco inferiore al 17% (è il caso del primo quartile). Anche nel caso di dati reali lo stimatore CD risulta molto preciso, indifferentemente dal predittore usato tra il modello lineare ad effetti misti e il modello di regressione M-quantile[21]. Per quanto riguarda un confronto tra lo stimatore della funzione di ripartizione CD e lo stimatore della funzione di ripartizione RKM si rimanda al lavoro di Tzavidis *e altri* (2008b).

[21] Vista la situazione particolare per cui si è utilizzato il campione LSMS Albania del 2003 come popolazione di riferimento per la simulazione basata su disegno, si preferisce parlare di dati *pseudo*-reali.

Capitolo 4

Proposta di una stima per l'errore quadratico medio per lo stimatore Chambers-Dunstan della funzione di ripartizione per piccola area

4.1 Introduzione

In questo capitolo sarà presentata una proposta per la stima dell'errore quadratico medio dello stimatore della funzione di ripartizione nell'ambito della stima per piccole aree. Lo stimatore della funzione di ripartizione a cui si farà riferimento è quello proposto da Chambers e Dunstan (1986), e nella fattispecie della stima per piccole aree si farà riferimento allo stimatore della funzione di ripartizione CD che utilizza il modello di regressione M-quantile (3.27). La proposta per la stime dell'errore quadratico medio dello stimatore della funzione di ripartizione per piccola area è ottenuta basandosi principalmente sull'articolo di Lombardia *e altri* (2003), che è stato opportunamente adattato al problema della stima per piccole aree. Sarà proposto anche uno stimatore dell'errore quadratico medio dello stimatore dei quantili per piccola area, dove lo stimatore dei quantili per piccola area è ottenuto dallo stimatore della funzione di ripartizione CD per piccola area basato sul modello di regressione M-quantile (come modello di superpopolazione).

Per mostrare le performance dello stimatore proposto saranno presentate due simulazioni, una basata su modello (denominata anche simulazione model-based) e una basata su disegno (denominata anche simulazione design-based). Nell'ultima parte del capitolo sarà presentata la stima per provincia dei quartili del reddito equivalente disponibile in toscana ed il relativo errore quadratico medio, dove la provincia rappresenta la piccola area.

4.2 Stimatore bootstrap per l'errore quadratico medio dello stimatore Chambers-Dunstan per la funzione di ripartizione

4.2.1 Il metodo bootstrap: una breve introduzione

Il metodo bootstrap è stato presentato nella letteratura specializzata da Efron (1979). Sin dalla sua introduzione questa metodologia è stata applicata ad una vastissima gamma di problemi di stima nei

più disparati campi di studio. La letteratura inerente la metodologia bootstrap è vastissima e persino nei libri più noti, come Davison e Hinkley (1997), Shao e Tu (1995), Efron e Tibshirani (1993) o Hall (1992), non si tratta che alcuni aspetti di questa metodologia. In questo si vuole dare al lettore una minima panoramica della metodologia bootstrap in modo da inquadrare adeguatamente il tipo di stimatore che sarà proposto nei paragrafi successivi.

Nella maggior parte dei problemi statistici l'obiettivo è quello di conoscere il vero valore di un certo parametro, ϑ, riferito al carattere di interesse X di una popolazione tramite l'estrazione di un campione casuale, $s = \{1, \ldots, n\}$. Dal campione s si possono ottenere informazioni sul carattere di interesse X, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Utilizzando i dati campionari, $\boldsymbol{x}$, si può fare inferenza su ϑ tramite uno stimatore $T(\boldsymbol{x})$ per ϑ. Con il metodo bootstrap si vuole ottenere informazioni sulla relazione esistente tra ϑ e lo stimatore $T(\boldsymbol{X})$ utilizzando la relazione che esiste tra $T(\boldsymbol{x})$ e $T(\boldsymbol{x}*)$, dove $\boldsymbol{x}*$ è un vettore di osservazioni ottenuto campionando dai dati campionari $\boldsymbol{x}$. $\boldsymbol{x}*$ può essere ottenuto campionando con ripetizione dai dati campionari $\boldsymbol{x}$, ed in questo caso si tratta di bootstrap non parametrico, oppure campionando da una funzione di ripartizione parametrizzata dalla statistica $T(\boldsymbol{x})$, in questo caso si tratta di bootstrap parametrico. Spesso la statistica $T(\boldsymbol{x})$ è l'errore quadratico medio di un certo stimatore, su cui si vuole costruire un intervallo di confidenza[1].

Nell'approccio non parametrico non si fanno assunzioni sulla distribuzione dei dati o dei parametri. Si ipotizzi di conoscere un campione s di n osservazioni indipendenti sul carattere X, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Si consideri lo stimatore $T(\boldsymbol{x})$ per il parametro ϑ. Se si vuole usare il metodo bootstrap per ottenere un intervallo di confidenza per lo stimatore $T(\boldsymbol{x})$ oppure per ottenere una stima dell'errore quadratico medio dello stimatore $T(\boldsymbol{x})$, la procedura da seguire è la seguente:

1. Estrarre un campione casuale, $s*$, con ripetizione di n elementi dai dati campionari $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Il campione $s*$, denominato campione bootstrap, è formato dagli elementi $\boldsymbol{x}* = \{x_1*, \ldots, x_n*\}$.

2. Calcolare la statistica T sul campione bootstrap $s*$: $T = T(\boldsymbol{x}*)$.

3. Ripetere i passaggi 1 e 2 B volte, in questo modo si ottiene una stima della distribuzione bootstrap della statistica T.

L'intervallo di confidenza ad un livello di fiducia $1 - \alpha$ per $T(\boldsymbol{x})$ può essere ottenuto scegliendo i percentili $\alpha/2$ e $1 - \alpha/2$ della distribuzione bootstrap, ottenuta come descritto nel punto 3. In alternativa si può ottenere un intervallo di confidenza con un livello di fiducia $1 - \alpha$ utilizzando l'approssimazione normale:

$$T(\boldsymbol{x}) \pm z_{\alpha/2} \sqrt{\widehat{MSE}[T(\boldsymbol{x})]},$$

dove $\widehat{MSE}[T(\boldsymbol{x})]$ è la stima dell'errore quadratico medio dello stimatore $T(\boldsymbol{x})$ ottenuta dalla distribuzione bootstrap e $z_{\alpha/2}$ è l'$\alpha/2$-esimo percentile della distribuzione normale standard. $\widehat{MSE}[T(\boldsymbol{x})]$ si ottiene dalla somma della stima della varianza di $T(\boldsymbol{x})$ con la stima della distorsione di $T(\boldsymbol{x})$:

[1] Si ricorda che lo stimatore è una variabile casuale.

$$\hat{V}[T(\boldsymbol{x})] = B^{-1}\sum_{b=1}^{B} n^{-1}\sum_{i=1}^{n}(T(\boldsymbol{x}^*_{bi} - T(\boldsymbol{x})))^2$$
$$\widehat{Bias}[T(\boldsymbol{x})] = B^{-1}\sum_{b=1}^{B} n^{-1}\sum_{i=1}^{n}(T(\boldsymbol{x}^*_{bi} - T(\boldsymbol{x})))$$
$$\widehat{MSE}[T(\boldsymbol{x})] = \widehat{Bias}[T(\boldsymbol{x})]^2 + \hat{V}[T(\boldsymbol{x})].$$

Esistono in letteratura alcune sfumature rispetto alla metodologia presentata, ed esistono svariati metodi per il calcolo dell'intervallo di confidenza su uno stimatore o la stima dell'errore quadratico medio di un certo stimatore, tuttavia l'idea alla base dell'approccio bootstrap non parametrico è quella presentata.

Nell'approccio parametrico si assume che i dati seguano una certa distribuzione, $F_X(\boldsymbol{x};\vartheta)$, caratterizzata dal parametro incognito ϑ. Si ipotizzi di conoscere un campione s di n osservazioni indipendenti sul carattere X, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Si consideri lo stimatore $T(\boldsymbol{x})$ per il parametro ϑ. Se si vuole usare il metodo bootstrap per ottenere un intervallo di confidenza per lo stimatore $T(\boldsymbol{x})$ oppure per ottenere una stima dell'errore quadratico medio dello stimatore $T(\boldsymbol{x})$ allora la procedura da seguire è la seguente:

1. Sia $T(\boldsymbol{x})$ la stima di ϑ calcolata sul campione s. Estrarre un campione casuale, $s*$, di n unità dalla distribuzione $F_X(\cdot; T(\boldsymbol{x}))$. Il campione $s*$ è denominato campione bootstrap ed è formato dagli elementi $\boldsymbol{x}* = \{x_1*, \ldots, x_n*\}$ generati dalla distribuzione $F_X(\cdot; T(\boldsymbol{x}))$.

2. Calcolare la statistica T sul campione bootstrap $s*$: $T = T(\boldsymbol{x}*)$.

3. Ripetere i passi 1 e 2 B volte per ottenere una stima della distribuzione bootstrap parametrica della statistica T.

L'intervallo di confidenza o la stima dell'errore quadratico medio per lo stimatore $T(\boldsymbol{x})$ si ottengono esattamente come nel caso dell'approccio bootstrap non parametrico. L'unica differenza tra l'approccio parametrico e quello non parametrico sta nel fatto che il campione bootstrap nel caso parametrico è generato da una distribuzione nota (ad esempio normale, binomiale, gamma, beta, etc.) mentre nel caso non parametrico è generato campionando dai dati campionari.

Esiste un terzo approccio bootstrap, da alcuni denominato bootstrap semiparametrico (Carpenter e Bithell, 2000). In questo approccio si prevede che la variabile di interesse, Y, sia generata da un modello: $y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i$, dove $\boldsymbol{x}_i$ è il vettore di variabili ausiliarie riferite all'unità i-esima, $\boldsymbol{\beta}$ è un parametro incognito, $g()$ è la relazione che lega la variabile di interesse alle covariate e ε_i è l'errore del modello per l'unità i-esima. Si consideri lo stimatore $T(\boldsymbol{y})$ per il parametro ϑ. Si vuole usare il metodo bootstrap per ottenere un intervallo di confidenza per lo stimatore $T(\boldsymbol{y})$ oppure per ottenere una stima dell'errore quadratico medio dello stimatore $T(\boldsymbol{y})$. La procedura da seguire per questo approccio è la seguente:

1. Stimare il modello $y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i$.

2. Ottenere i residui del modello: $\epsilon_i = y_i - g(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$, dove $\hat{\boldsymbol{\beta}}$ è una stima di $\boldsymbol{\beta}$.

3. Centrare i residui ϵ_i, in questo modo i residui hanno media 0. Si indichi con $\tilde{\epsilon}_i$ i residui centrati.

4. Estrarre un campione casuale $s*$ di n elementi dai residui centrati, $\tilde{\epsilon}_i$. $s*$ è il campione bootstrap e i residui campionati sono denominati residui bootstrap, $\{\tilde{\epsilon}_1^*, \ldots, \tilde{\epsilon}_n^*\}$.

5. Generare i dati $\boldsymbol{y}*$ dal modello stimato e dai residui bootstrap: $y_i^* = g(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) + \tilde{\epsilon}_i^*$.

6. Calcolare la statistica T sui dati bootstrap $\boldsymbol{y}*$: $T = T(\boldsymbol{y}*)$.

7. Ripetere i passi da 2 a 6 per B volte per ottenere una stima della distribuzione bootstrap semiparametrica della statistica T.

L'intervallo di confidenza o la stima dell'errore quadratico medio per lo stimatore $T(\boldsymbol{y})$ si ottengono come mostrato nel caso dell'approccio bootstrap non parametrico. Esistono molte varianti per l'approccio bootstrap semiparametrico. Una di queste sarà utilizzata per la stima dell'errore quadratico medio della funzione di ripartizione nel proseguo di questo capitolo.

Uno dei problemi principali dell'approccio bootstrap è la scelta di B. Alcuni autori (Efron e Tibshirani, 1993; Davison e Hinkley, 1997) consigliano di porre B tra 1000 e 2000. Il numero di replicazioni bootstrap dipende dal tipo di problema e dal tipo di approccio bootstrap che si vuole utilizzare. Una regola universale non sembra esistere al momento e il continuo aumento della potenza di calcolo a basso costo rende il problema sempre meno importante. Tuttavia nei casi in cui la stima della statistica $T(\boldsymbol{x})$ è molto lenta, anche per i computer odierni, si è costretti a usare poche replicazioni bootstrap, ma in genere mai meno di 400 (o 399)[2]. Un suggerimento per ovviare alla lentezza della procedura bootstrap, nei casi ove essa si presenti, è quello di utilizzare software di programmazione di basso livello, come FORTRAN o C.

La stima bootstrap non deve essere intesa come un metodo di stima "empirico", dove empirico sta a indicare che il metodo funziona frequentemente e quindi può essere usato. A partire da Efron (1979) sono state proposte diverse dimostrazioni teoriche sulla validità del metodo bootstrap. Bickel e Freedman (1981) dimostrano che la procedura bootstrap per la stima di parametri e intervalli di confidenza è asintoticamente corretta in diverse situazioni, come la stima dei quntili, delle statistiche t e nelle funzioni di von Mises.

4.2.2 Lo stimatore bootstrap per l'errore quadratico medio dello stimatore Chambers-Dunstan per la funzione di ripartizione: una proposta di Lombardia *e altri* (2003)

Lombardia *e altri* (2003) propongono un approccio bootstrap per la stima dell'errore quadratico medio dello stimatore della funzione di ripartizione proposto da Chambers e Dunstan (1986). La proposta

[2]Usare numeri dispari nelle replicazioni bootstrap è molto comodo nel caso in cui si voglia ottenere intervalli di confidenza tramite i percentili, infatti in una sequenza, sufficientemente lunga, di numeri dispari i percentili sono individuati senza bisogno di interpolazione.

di Lombardia *e altri* (2003) si basa sul modello di regressione lineare semplice come modello di superpopolazione nonostante Chambers *e altri* (1992) avessero già proposto uno stimatore dell'errore quadratico medio dello stimatore CD della funzione di ripartizione asintoticamente corretto in questa circostanza. Tuttavia l'approccio bootstrap è facilmente adattabile ad altri tipi di modelli di superpopolazione, ed anche nel caso di modello di superpopolazione lineare semplice è più gestibile del complesso stimatore dell'errore quadratico medio proposto da Chambers *e altri* (1992).

Si consideri una popolazione finita Ω di N unità, e siano $\{y_1, \ldots, y_N\}$ e $\{x_1, \ldots, x_N\}$ i valori rispettivamente della variabile di interesse Y e della variabile ausiliaria X misurati sulle N unità di Ω. Si consideri, inoltre, il campione casuale semplice s di $n \leq N$ unità, estratto senza ripetizione da Ω. La variabile ausiliaria X si considera nota senza errore per tutte le unità della popolazione Ω mentre Y, la variabile di interesse, è nota solo per le unità campionate, $\{y_1, \ldots, y_n\}$[3]. Si consideri il seguente modello di superpopolazione per la variabile di interesse Y:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{4.1}$$

dove α e β sono parametri incogniti e non osservabili e le ε_i sono variabili casuali indipendenti e identicamente distribuite con media zero e funzione di ripartizione G, $G(t) = Pr(\varepsilon \leq t)$. Si ricordi che secondo le ipotesi fatte lo stimatore CD per la funzione di ripartizione è:

$$\hat{F}_N(t) = N^{-1}\left[\sum_{i\in s} I(y_i \leq t) + \sum_{k\in r} n^{-1} \sum_{i\in s} I(\hat{\alpha} + \hat{\beta}x_k + \epsilon_i \leq t)\right],$$

dove $\hat{\alpha}$ e $\hat{\beta}$ sono le stime OLS dei coefficienti di regressione α e β del modello di superpopolazione (4.1) e ϵ_i è il residuo del modello di regressione (4.1), $\epsilon_i = y_i - \hat{\alpha} - \hat{\beta}x_i$.

Per la procedura bootstrap proposta da Lombardia *e altri* (2003) è necessario stimare la funzione di ripartizione G dell'errore ε nel modello di regressione (4.1). Gli autori suggeriscono due approcci per la stima di G. Il primo consiste nella stima G tramite la funzione di ripartizione empirica dei residui di regressione (ϵ) centrati:

$$\hat{G}_N(t) = G_n(t) = n^{-1} \sum_{i\in s} I(\epsilon_i - \bar{\epsilon} \leq t),$$

dove $\bar{\epsilon}$ è la media aritmetica dei residui di regressione: $\bar{\epsilon} = n^{-1}\sum_{i\in s} \epsilon_i$. Si farà riferimento a questo approccio con la denominazione di *distribuzione empirica* (dei residui) e sarà indicato con $\hat{G}^e_N(t)$.

Nel secondo approccio la stima della funzione di ripartizione G degli errori di regressione è ottenuta con uno stimatore per la funzione di ripartizione di tipo non parametrico:

$$\hat{G}_N(t) = n^{-1} \sum i \in sK\left(\frac{t - (\epsilon_i - \bar{\epsilon})}{h}\right),$$

dove h è il parametro di smoothing e K è la funzione di ripartizione ottenuta da una funzione di densità simmetrica e limitata[4]:

$$K(t) = \int_{-\infty}^{t} k(z)dz.$$

Si farà riferimento a questo approccio con la denominazione di *distribuzione smooth* (dei residui), indicato con la notazione $\hat{G}^h_N(t)$.

[3] In questo capitolo si considera N noto.

[4] Nel senso che $k(t) = dK(t)/dt$.

Tramite la stima $\hat{G}_N(t)$ di $G_N(t)$ si costruisce una popolazione (finita) bootstrap[5], Ω^*, formata dai caratteri Y^* e X (carattere noto per tutte le unità della popolazione in modo ceto e senza errore), dove $Y^* = \{y_1^*, \ldots, y_N^*\}$ è generato conformemente al modello

$$y_i^* = \hat{\alpha} + \hat{\beta} x_i + \varepsilon_i^*, \tag{4.2}$$

dove $\hat{\alpha}$ e $\hat{\beta}$ sono le stime OLS del modello (4.1) ottenute dal campione s e ε_i^* sono residui campionati dalla funzione di ripartizione $\hat{G}_N$.

La funzione di ripartizione della variabile Y^* per la popolazione bootstrap Ω^* è nota ed è:

$$F_N^*(t) = N^{-1} \sum_{i \in \Omega^*} I(y_i^* \leq t). \tag{4.3}$$

Si estragga un campione casuale semplice (senza reintroduzione), s^*, di $n \leq N$ unità dalla popolazione bootstrap Ω^*. Si consideri le variabili Y^* e X associate alle unità del campione s^*, ovvero $\{y_1^*, \ldots, y_n^*\}$ e $\{x_1, \ldots, x_n\}$. Dai valori $\{y_1^*, \ldots, y_n^*\}$ e $\{x_1, \ldots, x_n\}$, e considerando valido il modello di superpopolazione (4.2), si ottiene una stima della funzione di ripartizione per la variabile Y^* utilizzando lo stimatore CD:

$$\hat{F}_N^*(t) = N^{-1} \left[\sum_{i \in s^*} I(y_i^* \leq t) + \sum_{k \in r^*} n^{-1} \sum_{i \in s^*} I(\hat{\alpha}^* + \hat{\beta}^* x_k + \epsilon_i^* \leq t) \right], \tag{4.4}$$

dove $\hat{\alpha}^*$ e $\hat{\beta}^*$ sono le stime OLS del modello (4.2) ottenute con i dati del campione s^*, ϵ_i^* è il residuo del modello di regressione (4.2), $\epsilon_i^* = y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i$ e r^* indica l'insieme delle unità della popolazione bootstrap non campionate, $r^* = \Omega^* - s^*$.

Lombardia *e altri* (2003) dimostrano che la funzione di ripartizione della popolazione bootstrap Ω^*, $F_N^*(t)$ (4.3), e la sua stima ottenuta con lo stimatore CD, utilizzando l'approccio smooth, $\hat{F}_N^*(t)$ (4.4), hanno la stessa varianza e lo stessa distorsione (sarebbe più corretto dire la stessa non distorsione) asintotica dello stimatore della funzione di ripartizione CD, $\hat{F}_N(t)$, rispetto alla vera funzione di ripartizione della popolazione finita Ω, $F_N(t)$ (si consulti Shao e Tu (1995), pagina 76, per un approfondimento sulla validità della metodologia bootstrap).

Si consideri le seguenti condizioni di regolarità:

C-i. La funzione di densità k per la stima kernel della distribuzione dell'errore del modello di superpopolazione deve essere simmetrica, limitata, derivabile almeno due volte e con un momento finito almeno di ordine 2. La derivata prima e seconda di k devono essere limitate, uniformemente continue e devono valere le seguenti condizioni:

$$\begin{aligned} \int |t(\log t)|^{1/2} |k^{(l)} dt| &< \infty \quad (l = 0, 1, 2), \\ \int_0^1 (\log t^{-1})^{1/2} dV_l(t)^{1/2} &< \infty, \end{aligned}$$

dove $k^{(l)}$ indica la derivata di ordine l della funzione di densità k, V è il modulo di continuità di $k^{(l)}$, $(l = 1, 2)$. Inoltre la trasformata di Fourier di k non deve essere uguale ad uno in nessun intorno dello zero.

[5]Dove $\hat{G}_N$ indica indifferentemente $\hat{G}_N^e$ o $\hat{G}_N^h$.

C-ii. Il parametro di bandwidth, h, nella stima kernel della distribuzione dell'errore del modello di superpopolazione deve valere $h = n^{-\gamma}$, con $\gamma < 1/8$.

C-iii. La funzione di densità, g, dell'errore di regressione nel modello di superpopolazione deve essere limitata e derivabile almeno due volte, e la derivata prima e seconda devono essere anch'esse limitate e uniformemente continue.

C-iv. Per n ed N che tendono ad infinito, il rapporto n/N deve tendere ad una costante, π, con valori compresi tra zero e uno:

$$\lim_{n\to\infty, N\to\infty} \frac{n}{N} = \pi \quad \pi \in [0, 1].$$

C-v. $\{x_1, \ldots, x_N\}$, i valori della variabile ausiliaria X per tutta la popolazione, appartengono ad un intervallo finito. La variabile ausiliaria per le unità campionate, $x_i, i \in s$, e per le unità non campionate, $x_k, k \in r$, sono generate dalla stessa funzione di densità δ e vale l'approssimazione

$$n^{-1} \sum_{i\in s} I(x_i \leq t) \to \int_{-\infty}^{t} \delta(u)du, \quad (N-n)^{-1} \sum k \in rI(x_k \leq t) \to \int_{-\infty}^{t} \delta(u)du.$$

C-vi. Per l'errore del modello di superpopolazione, ε, deve valere: $E[|\varepsilon|^{2+\gamma}] < \infty$, con $\gamma > 0$.

Usando queste condizioni di regolarità Lombardia *e altri* (2003) dimostrano i seguenti teoremi:

Teorema 4.2.1 *Date le condizioni di regolarità* ***C-i.-C-v.****, allora*

$$E_*[\hat{F}_N^*(t) - F_N^*(t)] = O(n^-1).$$

Teorema 4.2.2 *Date le condizioni di regolarità* ***C-i.-C-v.*** *e le seguenti definizioni:*

i. $\hat{g}^h(t) = d\hat{G}_N^h(t)/dt$*, dove* $\hat{g}^h(t)$ *è la stima kernel della densità dell'errore del modello* (4.1) *e* $G_N^h(t)$ *ne è la stima smooth della funzione di ripartizione.*

ii. $\hat{\sigma}^2 = n^{-1} \sum_{i\in s} (\epsilon_i - \bar{\epsilon})^2$*, la varianza dei residui del modello di regressione* (4.1)*.*

iii. $\mu = \int x\delta(x)dx$*, la media della variabile ausiliaria.*

iv. $\varsigma^2 = \left(\int x^2\delta(x)dx\right) - \mu^2$*, la varianza asintotica della variabile ausiliaria.*

v. $(x \wedge y) = \min(x, y)$.

vi. $\pi = n/N$,
allora

$$
\begin{aligned}
V[\hat{F}^*_N(t) - F^*_N(t)] = & \\
&= n^{-1}(1-\pi)^2 \left\{ \varsigma^{-2}\hat{\sigma}^2 \left(\int (x-\mu)\hat{g}^h(t-\hat{\alpha}-\hat{\beta}x)\delta(x)dx \right)^2 + \right. \\
&+ \iint \hat{G}^h_N[(t-\hat{\alpha}-\hat{\beta}x) \wedge (t-\hat{\alpha}-\hat{\beta}y)]\delta(x)\delta(y)dxdy - \\
&\left. - \left(\int \hat{G}^h_N(t-\hat{\alpha}-\hat{\beta}x)\delta(x)dx \right)^2 \right\} + \\
&+ N^{-1}(1-\pi) \int [\hat{G}^h_N(t-\hat{\alpha}-\hat{\beta}x) - \hat{G}^h_N(t-\hat{\alpha}-\hat{\beta}x)^2]\delta(x)dx + o(n^-1).
\end{aligned}
$$

L'asterisco nel valore atteso del teorema 4.2.1, E_*, indica che il valore atteso è fatto rispetto al modello con cui si genera la popolazione bootstrap (4.2).

Lombardia *e altri* (2003) dimostrano anche la convergenza in distribuzione alla normale standard dello stimatore della funzione di ripartizione CD e della sua versione bootstrap:

Teorema 4.2.3 *Date le condizioni **iii.-vi.**, allora*

$$\frac{\hat{F}_N(t) - F_N(t)}{Var[\hat{F}_N(t) - F_N(t)]^{1/2}} \xrightarrow{d} N(0,1).$$

Teorema 4.2.4 *Date le condizioni **i.-vi.**, allora*

$$\frac{\hat{F}^*_N(t) - F^*_N(t)}{Var[\hat{F}^*_N(t) - F^*_N(t)]^{1/2}} \xrightarrow{d} N(0,1).$$

Considerati i modelli di superpopolazione (4.1) e (4.2) Lombardia *e altri* (2003) dimostrano che la stima bootstrap della distorsione e della varianza dello stimatore della funzione di ripartizione CD sono rispettivamente:

$$E_*[\hat{F}^*_N(t) - F^*_N(t)] = (Nn)^{-1} \sum_{k \in r^*} \sum_{i \in s^*} E_*[\hat{G}^h_N(\hat{t}_k - (\hat{\beta}^* - \hat{\beta})(x_k - x_i)) - \hat{G}^h_N(\hat{t}_k) + O(n^{-1})], \quad (4.5)$$

e

$$
\begin{aligned}
V_*[\hat{F}^*_N(t) - F^*_N(t)] = & \\
&= N^2 \sum_{k \in r^*} [\hat{G}^h_N(\hat{t}_k) - \hat{G}^h_N(\hat{t}_k)^2] + \\
&+ (nN^2)^{-1} \sum_{k_1, k_2 \in r^*} [\hat{G}^h_N(\hat{t}_{k_1} \wedge \hat{t}_{k_2}) - \hat{G}^h_N(\hat{t}_{k_1})\hat{G}^h_N(\hat{t}_{k_2})] + \qquad (4.6) \\
&+ (nN^2)^{-1}\hat{\sigma}^2\hat{\sigma}^{*-2}_x \left[\sum_{k \in r^*} (x_k - \bar{x})\hat{g}^h(\hat{t}_k) \right]^2 + O(n^{-1}),
\end{aligned}
$$

dove l'asterisco nel valore atteso, E_* (4.5) e nella varianza attesa (4.6), V_*, indica che il valore atteso è fatto rispetto al modello con cui si genera la popolazione bootstrap (4.2), $\hat{t}_k = t - \hat{\alpha} - \hat{\beta}x_k$, $\hat{\sigma}_x^{*2} = n^{-1}\sum_{i\in s^*}(x_i - \bar{x})^2$ è la varianza campionaria della X nel campione bootstrap s^*, $\hat{\alpha}$ e $\hat{\beta}$ sono le stime OLS dei coefficienti di regressione del modello (4.1) e $\hat{\alpha}^*$ e $\hat{\beta}^*$ sono le stime OLS dei coefficienti di regressione del modello (4.2); le altre grandezze sono quelle elencate nei punti **i.-vi.** del teorema 4.2.2.

Se il modello di superpopolazione non è il modello di regressione lineare (4.1) non si conosce la forma della (4.5) e (4.6). Lombardia *e altri* (2003) propongono una simulazione Monte Carlo per ottenere un'approssimazione della (4.5) e (4.6).

Si consideri una popolazione finita Ω di N unità, e siano $\{y_1, \ldots, y_N\}$ e $\{x_1, \ldots, x_N\}$ i valori rispettivamente della variabile di interesse Y e della variabile ausiliaria X misurati sulle N unità di Ω. Si consideri, inoltre, il campione casuale semplice s di $n \leq N$ unità, estratto senza ripetizione da Ω. La variabile ausiliaria X si considera nota senza errore per tutte le unità della popolazione Ω mentre Y, la variabile di interesse è nota solo per le unità campionate, $\{y_1, \ldots, y_n\}$. Si consideri il modello di superpopolazione (4.1) e la popolazione bootstrap Ω^* generata, come descritto precedentemente, dal modello (4.2) utilizzando indifferentemente uno dei due approcci, empirico o smooth, per la stima di $G_N(t)$.

La simulazione Monte Carlo per la stima bootstrap dell'errore quadratico medio dello stimatore CD della funzione di ripartizione consiste nel generare B popolazioni bootstrap, $\Omega^{*b}, b = (1, \ldots, B)$, secondo lo schema descritto precedentemente ed estrarre da ognuna delle B popolazione bootstrap L campioni casuali semplici senza reintroduzione, s^{*l}, in modo da rispettare la frazione di campionamento del campione "originario" s. L'approssimazione Monte Carlo degli stimatori bootstrap della distorsione e della varianza dello stimatore della funzione di ripartizione CD sono rispettivamente:

$$\widehat{Bias}[\hat{F}_N^*(t) - F_N^*(t)] = B^{-1}\sum_{b=1}^{B} L^{-1}\sum_{l=1}^{L}(\hat{F}_N^{*bl}(t) - F_N^{*b}(t)),$$

e

$$\hat{V}[\hat{F}_N^*(t) - F_N^*(t)] = B^{-1}\sum_{b=1}^{B} L^{-1}\sum_{l=1}^{L}(\hat{F}_N^{*bl}(t) - \bar{\hat{F}}_N^{*b}(t))^2,$$

dove $F_N^{*b}(t)$ è la funzione di ripartizione della b-esima popolazione bootstrap, $\hat{F}_N^{*bl}(t)$ è la stima CD per la funzione di ripartizione $F_N^{*b}(t)$ della b-esima popolazione bootstrap ottenuta dal campione l-esimo, s^{*l}, estratto dalla b-esima popolazione bootstrap e $\bar{\hat{F}}_N^{*b}(t)$ è la media delle L stime CD, $\hat{F}_N^{*bl}$, ottenute dagli L campioni estratti dalla b-esima popolazione bootstrap: $\bar{\hat{F}}_N^{*b}(t) = L^{-1}\sum_{l=1}^{L}\hat{F}_N^{*bl}$. L'approssimazione bootstrap dell'errore quadratico medio dello stimatore CD della funzione di ripartizione è:

$$\widehat{MSE}[\hat{F}_N] \approx \widehat{MSE}[\hat{F}_N^*] = \hat{V}[\hat{F}_N^*(t) - F_N^*(t)] + \widehat{Bias}[\hat{F}_N^*(t) - F_N^*(t)]^2,$$

con un'approssimazione sempre più precisa al crescere di B ed L.

Per l'applicazione di questo metodo a casi reali Lombardia *e altri* (2003) consigliano di porre $B = 50$ e $L = 100$. Per quanto riguarda la costruzione di un intervallo di confidenza ad un livello di fiducia di $1 - \alpha$ per la funzione di ripartizione stimata con il metodo CD, Lombardia *e altri* (2003) consigliano di utilizzare l'approssimazione normale: $[\hat{F}_N(t) \pm z_{\alpha/2}MSE[\hat{F}_N^*]^{1/2}]$, dove $z_{\alpha/2}$ è il percentile $\alpha/2$ della distribuzione normale standard. L'approccio alla costruzione dell'intervallo di

confidenza con l'uso dei percentili è altresì considerato nell'articolo di Lombardia *e altri* (2003), ma gli autori rimandano a lavori futuri per un approfondimento su questo tema.

4.3 Stimatore bootstrap per l'errore quadratico medio dello stimatore Chambers-Dunstan per la funzione di ripartizione per piccola area

Sia $\Omega = \{1, \ldots, N\}$ una popolazione finita e $\boldsymbol{y} = [y_1, \ldots, y_N]'$ il vettore della variabile d'interesse (per tutte le unità della popolazione Ω). Si consideri un campione casuale $s \in \Omega$, di $n \leq N$ unità. Sia $r = \Omega - s$ l'insieme delle unità non campionate tale che $\boldsymbol{y} = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, dove $\boldsymbol{y}_s$ è il vettore noto delle n unità campionate e $\boldsymbol{y}_r$ il vettore non noto delle $N - n$ unità non campionate. Si consideri le unità della popolazione appartenenti a m gruppi o aree. Sia $\boldsymbol{y}_{s_i}$ il vettore delle unità campionate nell'area $i = (1, \ldots, m)$ e $\boldsymbol{y}_{r_i}$ il vettore delle unità non campionate nell'area i, dove $s_i = \{1, \ldots, n_i\}$, $s = \{s_1, \ldots, s_m\}$, $\sum_{i=1}^{m} n_i = n$, $r = \{r_1, \ldots, r_m\}$ e $\sum_{i=1}^{m} N_i - n_i = N - n$. Si ipotizzi che le variabili ausiliarie $\{X_1, \ldots, X_p\}$ siano note per tutte le unità della popolazione Ω in modo certo e senza errore, e sia $\boldsymbol{x}_{ji} = [x_{j1}, \ldots, x_{jp}]'$ il vettore delle variabili ausiliarie relativo all'unità j-esima appartenente all'area i.

Si ipotizzi che la variabile di interesse Y sia generata dal seguente modello di superpopolazione:

$$y_{ji} = \boldsymbol{x}_{ji}\boldsymbol{\beta}_{\psi}(\theta_i) + \varepsilon_{ji}, \tag{4.7}$$

che è il modello di regressione M-quantile presentato nel capitolo 2.

Lo stimatore bootstrap per la stima dell'errore quadratico medio dello stimatore CD proposto da Lombardia *e altri* (2003) si può facilmente adattare al contesto della stima per piccole aree. Utilizzando il modello di superpopolazione (4.7) si ottiene, come mostrato nel capitolo 3, una stima della funzione di ripartizione per piccola area utilizzando lo stimatore proposto da Chambers e Dunstan (1986). Applicando il metodo bootstrap presentato nel paragrafo precedente allo stimatore della funzione di ripartizione CD per piccola area si ottiene la stima bootstrap dell'errore quadratico medio dello stimatore CD della funzione di ripartizione per piccola area.

Sia $G(t)$ la funzione di ripartizione dell'errore (ε_{ji}) del modello di regressione M-quantile (4.7) e sia $\epsilon_{ji} = y_{ji} - \boldsymbol{x}_{ji}\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_i)$ il residuo, dove $\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_j)$ è la stima del vettore dei coefficienti di regressione del modello di regressione M-quantile. Per la procedura bootstrap è necessario stimare la funzione di ripartizione $G(t)$ dell'errore ε nel modello di regressione (4.7). Nel caso della stima per piccole aree si suggeriscono quattro approcci per la stima di $G(t)$. Il primo consiste nella stima di $G(t)$ tramite la funzione di ripartizione empirica dei residui di regressione (ϵ) centrati di tutte le piccole aree:

$$\hat{G}(t) = G_n(t) = n^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} I(\epsilon_{ji} - \bar{\epsilon} \leq t), \tag{4.8}$$

dove $\bar{\epsilon}$ è la media aritmetica dei residui di regressione: $\bar{\epsilon} = n^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} \epsilon_{ji}$. Si farà riferimento a questo approccio con la denominazione di *distribuzione empirica non condizionata* (dei residui) e sarà indicato con $\hat{G}^e(t)$.

Il secondo approccio consiste nella stima di $G(t)$ utilizzando la funzione di ripartizione empirica dei residui di regressione centrati di una data area i:

$$\hat{G}_i(t) = G_{n,i}(t) = n_i^{-1} \sum_{j \in s_i} I(\epsilon_{ji} - \bar{\epsilon}_i \leq t), \tag{4.9}$$

dove $\bar{\epsilon}_i$ è la media aritmetica dei residui di regressione nell'area i: $\bar{\epsilon}_i = n_i^{-1} \sum_{j \in s_i} \epsilon_{ji}$. Si farà riferimento a questo approccio con la denominazione di *distribuzione empirica condizionata* (dei residui) e sarà indicato con $\hat{G}_i^e(t)$.

Nel terzo approccio la stima della funzione di ripartizione $G(t)$ degli errori di regressione è ottenuta con uno stimatore per la funzione di ripartizione di tipo non parametrico, in cui si considerano i residui di tutte le m piccole aree:

$$\hat{G}(t) = n^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} K\left(\frac{t - (\epsilon_{ji} - \bar{\epsilon})}{h}\right), \tag{4.10}$$

dove h è il parametro di smoothing e K è la funzione di ripartizione ottenuta da una funzione di densità simmetrica e limitata

$$K(t) = \int_{-\infty}^{t} k(z)dz.$$

Si farà riferimento a questo approccio con la denominazione di *distribuzione smooth non condizionata* (dei residui), indicato con la notazione $\hat{G}^h(t)$.

L'ultimo approccio per la stima della funzione di ripartizione $G(t)$ si basa sull'utilizzo dello stimatore non parametrico proposto nella (4.10) in cui si considerano solo i residui di una certa area i:

$$\hat{G}_i(t) = n_i^{-1} \sum_{j \in s_i} K\left(\frac{t - (\epsilon_{ji} - \bar{\epsilon}_i)}{h_i}\right), \tag{4.11}$$

dove h_i è il parametro di smoothing per la piccola area i e K è la stessa funzione di ripartizione presente nella (4.10). Si farà riferimento a questo approccio con la denominazione di *distribuzione smooth condizionata* (dei residui), indicato con la notazione $\hat{G}_i^h(t)$.

Nei risultati empirici che saranno presentati nei paragrafi successivi per K è stato usato, come suggerito da Lombardia *e altri* (2003), il kernel Epanechnikov:

$$k(t) = \frac{3}{4}(1 - t^2) I(|t| < 1),$$

mentre i parametri di smooth h e h_i sono determinati minimizzando una funzione di cross-validation secondo il criterio proposto da Bowman *e altri* (1998). Nel caso di approccio smooth non condizionato h è determinato minimizzando la seguente funzione cross-validation:

$$CV(h) = n^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} \int [I((\epsilon_{ji} - \bar{\epsilon}) \leq t) - \hat{G}_{-j}(t)]^2 dt,$$

dove $\hat{G}_{-j}(t)$ equivale alla stima $\hat{G}^h(t)$ in cui si omette l'unità j-esima. Per l'approccio smooth condizionato h_i è determinato dalla minimizzazione di:

$$CV(h_i) = n^{-1} \sum_{j \in s_i} \int [I((\epsilon_{ji} - \bar{\epsilon}_i) \leq t) - \hat{G}_{-ji}(t)]^2 dt,$$

dove $\hat{G}_{-ji}(t)$ equivale alla stima $\hat{G}_i^h(t)$ in cui si omette l'unità j-esima della piccola area i. Scegliere h e h_i con questo criterio è asintoticamente equivalente (si veda in proposito Li e Racine (2007), pagina 23) a scegliere h e h_i in termini di minimizzazione dell'errore quadratico medio integrato,

$IMSE[\hat{G}(t)] = \int[\hat{G}(t) - G(t)]$[6]. Negli studi di simulazione presenti nei paragrafi successivi le stime smooth in questione sono state ottenute utilizzando il pacchetto *np* (Tristen Hayfield and Jeffrey S. Racine (2008). Nonparametric Econometrics: The np Package. Journal of Statistical Software 27(5). URL http://www.jstatsoft.org/v27/i05/) nell'ambiente software R (R Development Core Team, 2008) che implementa le tecniche di stima appena illustrate.

Riassumendo, i quattro approcci per la stima della funzione di ripartizione dell'errore $G(t)$ del modello di superpopolazione sono:

i. Approccio empirico non condizionato: $\hat{G}^e(t)$ (4.8).

ii. Approccio empirico condizionato: $\hat{G}_i^e(t)$ (4.9).

iii. Approccio smooth non condizionato: $\hat{G}^h(t)$ (4.10).

iv. Approccio smooth condizionato: $\hat{G}_i^h(t)$ (4.11).

Utilizzando uno dei quattro approcci proposti per la stima di $G(t)$, sintetizzati dalla scrittura unica $\hat{G}(t)$, si costruisce una popolazione (finita) bootstrap, Ω^*, formata dai caratteri Y^* e $X_1, \ldots, X_p$, dove $X_1, \ldots, X_p$ sono caratteri noti per tutte le unità della popolazione in modo ceto e senza errore. $Y^* = \{\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_m^*\}$, con $\boldsymbol{y}_i = \{y_{1i}, \ldots, y_{n_i i}\}, i = (1, \ldots, m)$, è generato conformemente al modello di superpopolazione (4.7):

$$y_{ji}^* = \boldsymbol{x}_{ji}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + \varepsilon_{ji}^*, \tag{4.12}$$

dove $\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$ è la stima dei parametri ($\boldsymbol{\beta}_\psi$ e θ_i) del modello (4.7) e ε_i^* sono residui campionati dalla funzione di ripartizione $\hat{G}(t)$[7].

La funzione di ripartizione della variabile Y^* per l'area i della popolazione bootstrap Ω^* è nota senza errore, ed è:

$$F_{N,i}^*(t) = N_i^{-1} \sum_{j \in \Omega_i^*} I(y_{ji}^* \leq t) = N_i^{-1} \left[\sum_{j \in s_i} I(y_{ji}^* \leq t) + \sum_{k \in r_i} I(y_{ki}^* \leq t) \right].$$

Si estragga un campione casuale semplice (senza reintroduzione), s^*, di $n \leq N$ unità dalla popolazione bootstrap Ω^*, in modo da lasciare inalterata la struttura del campione originario s[8]. Si consideri le variabili Y^* e $X_1, \ldots, X_p$ associate alle unità del campione s^*, ovvero $\boldsymbol{y}_{ji}^*$ e $\boldsymbol{x}_{ji}$, $j \in s_i^*, i = (1, \ldots, m)$, dove s_i^* è il campione estratto da Ω^* nell'area i. Dai valori campionari $\boldsymbol{y}_{ji}^*$ e $\boldsymbol{x}_{ji}$, $j \in s_i^*, i = (1, \ldots, m)$ e considerando valido il modello di superpopolazione (4.12), si ottiene una stima della funzione di ripartizione per la variabile Y^* nell'area i utilizzando lo stimatore CD:

$$\hat{F}_{N,i}^*(t) = N_i^{-1} \left[\sum_{j \in s_i^*} I(y_{ji}^* \leq t) + \sum_{k \in r_i^*} n_i^{-1} \sum_{j \in s_i^*} I(\boldsymbol{x}_{ki}\hat{\boldsymbol{\beta}}_\psi^*(\hat{\theta}_i^*) + \epsilon_{ji}^* \leq t) \right],$$

[6] Questo è un criterio di ottimalità esteso per tutti i punti del dominio di $G(t)$.

[7] Ottenuta da uno dei quattro approcci proposti a pagina 72.

[8] In questo modo risulterà $n_i^* = n_i$, $N_i^* = N_i$, con $i = (1, \ldots, m)$.

dove $\hat{\boldsymbol{\beta}}^*_\psi(\hat{\theta}^*_i)$ è la stima dei parametri del modello (4.12) ottenuta con i dati del campione s^*, ϵ^*_{ji} è il residuo del modello di regressione (4.12), $\epsilon^*_{ji} = y^*_{ji} - \boldsymbol{x}_{ji}\hat{\boldsymbol{\beta}}^*_\psi(\hat{\theta}^*_i)$ e r^*_i indica l'insieme delle unità dell'area i nella popolazione bootstrap non campionate, $r^*_i = \Omega^*_i - s^*_i$.

La forma asintotica della distorsione e della varianza della stima bootstrap della funzione di ripartizione CD è nota solo se il modello di superpopolazione è il modello di regressione lineare semplice, (4.5) e (4.6). Se si ipotizzano modelli di superpopolazione diversi dal modello di regressione lineare si deve utilizzare l'approssimazione bootstrap per stimare distorsione e varianza dello stimatore della funzione di ripartizione CD, che per l'area i sono:

$$Bias_*[\hat{F}^*_{N,i}(t)] = E_*[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)],$$

e

$$V_*[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)] = E_*\left[((\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)) - E[(\hat{F}^*_{N,i}(t) - F^*_{N,i}(t))])^2\right]$$

dove l'asterisco nella varianza attesa e nella distorsione attesa, $Bias_*$ e Var_*, indica che il valore atteso è fatto rispetto al modello con cui si genera la popolazione bootstrap (4.12).

Si propone una simulazione Monte Carlo per ottenere un'approssimazione dello stimatore bootstrap di varianza e distorsione dello stimatore della funzione di ripartizione CD.

La simulazione Monte Carlo proposta consiste nel generare B popolazioni bootstrap, $\Omega^{*b}, b = (1, \ldots, B)$, secondo lo schema presentato precedentemente ed estrarre da ognuna delle B popolazione bootstrap L campioni casuali semplici senza reintroduzione s^{*l} in modo da rispettare la frazione di campionamento del campione "originario" s. L'approssimazione Monte Carlo degli stimatori bootstrap della distorsione e della varianza dello stimatore della funzione di ripartizione CD per l'area i sono rispettivamente:

$$\widehat{Bias}[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)] = B^{-1}\sum_{b=1}^{B} L^{-1}\sum_{l=1}^{L}(\hat{F}^{*bl}_{N,i}(t) - F^{*b}_{N,i}(t)),$$

e

$$\hat{V}[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)] = B^{-1}\sum_{b=1}^{B} L^{-1}\sum_{l=1}^{L}(\hat{F}^{*bl}_{N,i}(t) - \bar{\hat{F}}^{*b}_{N,i}(t))^2,$$

dove $F^{*b}_{N,i}(t)$ è la funzione di ripartizione dell'area i della b-esima popolazione bootstrap, $\hat{F}^{*bl}_{N,i}(t)$ è la stima CD per la funzione di ripartizione $F^{*b}_N(t)$ della b-esima popolazione bootstrap ottenuta dal campione l-esimo nell'area i, s^{*l}_i, estratto dalla b-esima popolazione bootstrap e $\bar{\hat{F}}^{*b}_{N,i}(t)$ è la media delle L stime CD nell'area i, $\hat{F}^{*bl}_{N,i}$, ottenute dagli L campioni estratti dalla b-esima popolazione bootstrap: $\bar{\hat{F}}^{*b}_{N,i}(t) = L^{-1}\sum_{l=1}^{L}\hat{F}^{*bl}_{N,i}$. L'approssimazione bootstrap dell'errore quadratico medio dello stimatore CD della funzione di ripartizione per l'area i è:

$$\widehat{MSE}[\hat{F}_{N,i}] \approx \hat{MSE}[\hat{F}^*_{N,i}] = \hat{V}[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)] + \widehat{Bias}[\hat{F}^*_{N,i}(t) - F^*_{N,i}(t)]^2, \tag{4.13}$$

con un'approssimazione sempre più precisa al crescere di B ed L. L'intervallo di confidenza per la stima della funzione di ripartizione CD si ottiene utilizzando l'approssimazione normale: $[\hat{F}_{N,i}(t) \pm z_{\alpha/2}\widehat{MSE}_i[\hat{F}^*_{N,i}]^{1/2}]$, dove $z_{\alpha/2}$ è il percentile $\alpha/2$ della distribuzione normale standard.

La stima dell'errore quadratico medio per i quantili ottenuti dalla stima della funzione di ripartizione CD per piccola area si ottiene semplicemente sostituendo nella procedura bootstrap presentata la funzione di ripartizione con il quantile, ovvero: $F_{N,i}(t) \rightarrow q_{Y,i}(\tau)$ e $\hat{F}_{N,i}(t) \rightarrow \hat{q}_{Y,i}(\tau)$, dove $q_{Y,i}(\tau)$ è il vero quantile di ordine τ della variabile di studio Y e $\hat{q}_{Y,i}(\tau)$ ne è la sua stima ottenuta secondo la (3.36)[9]. Per semplificare la notazione, d'ora in avanti si toglierà il pedice $_Y$ da $q_{Y,i}(\tau)$ e $\hat{q}_{Y,i}(\tau)$. La (4.13) nel caso di stima dei quantili, secondo quanto detto, diventa:

$$\widehat{MSE}[\hat{q}_i] \approx \widehat{MSE}[\hat{q}_i^*] = \hat{V}[\hat{q}_i^*(t) - q_i^*(t)] + \widehat{Bias}[\hat{q}_i^*(t) - q_i^*(t)]^2. \tag{4.14}$$

La procedura bootstrap presentata si può adattare facilmente alla stima dell'errore quadratico medio dello stimatore della media di piccola area basato sia sul modello di regressione M-quantile sia sul modello lineare ad effetti misti, per un approfondimento si consulti Tzavidis *e altri* (2008b). Inoltre lo schema presentato è facilmente generalizzabile per qualunque tipo di modello di superpopolazione, ad esempio si può sostituire il modello di regressione M-quantile con il modello lineare ad effetti misti nello stimatore CD della funzione di ripartizione ed ottenere, anche in questo caso, una stima dell'errore quadratico medio dello stimatore in questione.

Nei paragrafi successivi verrà verificato il comportamento dello stimatore dell'errore quadratico medio dello stimatore CD della funzione di ripartizione per piccola area basato sul modello di regressione M-quantile per diversi scenari e per i quattro approcci (pagina 72) per la stima della funzione di ripartizione dell'errore del modello di superpopolazione.

4.4 Simulazione Model-based per lo stimatore bootstrap dell'errore quadratico medio dello stimatore CD per la funzione di ripartizione per piccola area

Per verificare il comportamento dello stimatore dell'errore quadratico medio per i quantili ottenuti dallo stimatore CD della funzione di ripartizione per piccola area basato sul modello di regressione M-quantile (4.14), sono state generate due popolazioni, Ω_1 e Ω_2, di $N = 7430$ unità ciascuna, per simulare due diversi "scenari". La popolazione Ω_1 sarà creata con l'utilizzo di variabili casuali simmetriche, mentre la popolazione Ω_2 si baserà sull'utilizzo di variabili casuali asimmetriche, in questo modo si potrà osservare il comportamento degli stimatori proposti per questi due scenari "rappresentativi" di diverse realtà plausibili. Entrambe le popolazioni sono state divise in 30 aree ($m = 30$). Per entrambe le popolazioni sono state associate ad ogni unità due variabili, Y e X. La variabile Y è stata generata in funzione della variabile X con un modello lineare ad effetti misti del tipo:

$$y_{ji} = 1 + x_{ji} + u_i + \varepsilon_{ji},$$

dove y_{ji} e x_{ji} identificano rispettivamente la variabile di studio e la variabile ausiliaria per l'unità j-esima appartenente all'area i-esima, con $j = (1, \ldots, N_i)$ e $i = (1, \ldots, m)$. N_i indica la dimensione della popolazione nell'area i e per entrambe le popolazioni, Ω_1 e Ω_2, N_i varia tra 50 e 450. In Ω_1, x_{ji} è stata generata da una distribuzione normale con media μ_i e varianza 1. La media μ_i, fissa nella simulazione, varia tra 20 e 100. u_i è stato generato da una distribuzione normale con media 0 e varianza 1, mentre ε_{ji} da una distribuzione normale con media 0 e varianza 16. In Ω_2, x_{ji} è stata generata da una distribuzione χ^2 con gradi di libertà variabili tra 2 e 5 che sono stati tenuti fissi nella simulazione.

[9] In questo caso la stima della funzione di ripartizione CD per piccola area si basa sul modello M-quantile, ma la procedura per stimare l'errore quadratico medio per la stima dei quantili ottenuti dallo stimatore CD ha carattere generale e va bene per ogni modello di superpopolazione.

u_i è stato generato centrando una distribuzione χ^2 con 1 grado di libertà, mentre ε_{ji} è stato ottenuto centrando una distribuzione χ^2 con 3 gradi di libertà. Le popolazioni sono generate secondo i criteri esposti per ogni iterazione Monte Carlo. Dalle popolazioni Ω_1 e Ω_2 è stato estratto un campione casuale semplice, s. Il disegno di campionamento consiste nell'estrarre da ognuna delle 30 aree n_i unità, con $n_i = N_i/10$. Dal campione s si ottengono i valori campionari per la variabile di interesse Y in ogni area, $\boldsymbol{y}_i = \{y_{1i}, \ldots, y_{n_i i}\}$. La variabile X è considerata nota, in modo certo e senza errore, per tutte le unità della popolazione. Il campione s ha una dimensione totale di $n = \sum_{i=1}^{m} n_i = 743$ unità. N_i, N, n_i e n sono tenuti fissi nelle iterazioni della simulazione Monte Carlo. Tramite i dati campionari sono stati stimati, ad ogni iterazione Monte Carlo, i quartili della variabile di interesse Y, ottenuti dalla stima della funzione di ripartizione CD per piccola area basata sul modello di regressione M-quantile. Per ogni stima dei quartili di piccola area è stato stimato l'errore quadratico medio (4.14), secondo lo schema bootstrap presentato nel paragrafo precedente. Utilizzando l'approssimazione normale è stato calcolato un intervallo di confidenza al 95% per la stima dei quartili per piccola area, ottenuta secondo quanto detto precedentemente. Il focus in questa simulazione è la stima dell'errore quadratico medio della stima dei quantili per piccola area. D'ora in avanti si ometterà di specificare che i quantili stimati sono stati ottenuti dalla stima della funzione di ripartizione CD per piccola area utilizzando il modello di regressione M-quantile come modello di superpopolazione. La stima dell'errore quadratico medio per i quartili stimati della variabile Y per le popolazioni Ω_1 e Ω_2 è stata calcolata per ognuno dei quattro approcci alla stima della funzione di ripartizione dell'errore del modello di superpopolazione: empirico condizionato e non condizionato, e smooth condizionato e non condizionato (pagina 72). Vista la similitudine dei risultati dell'approccio condizionato e non condizionato sia nell'approccio empirico sia nell'approccio smooth si è deciso di riportare solo i risultati degli approcci non condizionati. Inoltre l'approccio condizionato, sia esso smooth o empirico, alla stima della funzione di ripartizione dell'errore del modello di superpopolazione è da utilizzarsi con molta cautela vista la scarsa numerosità di osservazioni che si ha, in genere, nelle diverse aree nell'ambito della stima per piccole aree. Si suggerisce di ricorrere all'approccio condizionato in caso di una numerosità sufficiente per una stima consistente della funzione di ripartizione dell'errore, sia essa parametrica o non parametrica, e per quei casi in cui la conoscenza a priori induca a ritenere una forte differenza nella distribuzione dell'errore del modello di superpopolazione tra le aree oggetto di studio. Si denoti con $\widehat{MSE}^{e,v}[\hat{q}_i(\tau)]$ la stima dell'errore quadratico medio della stima dei quartili per piccola area ottenuta con l'approccio empirico non condizionato nella procedura bootstrap proposta e con $\widehat{MSE}^{h,v}[\hat{q}_i(\tau)]$ quella ottenuta con l'approccio smooth non condizionato. Come termine di paragone per la stima dell'errore quadratico medio oggetto di studio si consideri la variabilità dello stimatore dei quartili di piccola area nella simulazione Monte Carlo, ovvero:

$$MSE[\hat{q}_i] = MC^{-1} \sum_{mc=1}^{MC} (\hat{q}_{i,mc} - q_i)^2 \tag{4.15}$$

dove $q_{i,mc}(\tau)$ è il vero valore del quantile di ordine τ nell'area i nell'iterazione mc della simulazione, $\hat{q}_{i,mc}(\tau)$ è la stima del quantile di ordine τ nell'area i nell'iterazione mc e MC è il numero di simulazioni Monte Carlo effettuate. L'errore quadratico medio della (4.15) è stato denominato per semplicità *vero* errore quadratico medio dello stimatore dei quantili di piccola area, dove per vero si intende l'approssimazione Monte Carlo dell'errore quadratico medio in questione; se $MC \to \infty$ allora l'errore quadratico medio della (4.15) sarebbe effettivamente il vero errore quadratico medio dello stimatore considerato.

La simulazione Monte Carlo è stata fatta ponendo $MC = 200$, mentre per quanto riguarda i parametri dello stimatore bootstrap è stato posto $L = 500$ e $B = 1$, dove L indica il numero di cam-

pioni estratti dalla popolazione artificiale bootstrap, $\Omega^{*b}, b = (1, \ldots, B)$, e B indica il numero delle popolazioni bootstrap generate. Purtroppo, vista la "lentezza" della procedura di stima dei quantili per piccola con il metodo CD, si è stati costretti a ridurre al minimo il numero delle iterazioni Monte Carlo e delle popolazioni bootstrap. Diverse prove hanno dimostrato una pari efficienza, misurata in termini di precisione della stima, dello stimatore bootstrap ottenuto con $B = 100$, $B = 50$ e $B = 1$. Per questo motivo si è posto $B = 1$, in questo modo è stato possibile porre $L = 500$ e $MC = 200$. Tuttavia sarebbe auspicabile poter effettuare la simulazione con $MC = 500$ e $B = 50$ in modo da avere un quadro più completo sul comportamento dello stimatore bootstrap per l'errore quadratico medio dei quantili per piccola area (o della funzione di ripartizione per piccola area).

Per le stime dell'errore quadratico medio della stima dei quartili per piccola area, ottenute con l'approccio empirico non condizionato e smooth non condizionato, è stato calcolato l'errore relativo, l'errore assoluto relativo e il tasso effettivo di copertura. L'errore relativo e l'errore assoluto relativo per la stima dell'errore quadratico medio di un dato quartile stimato, $\hat{q}_i(\tau)$, nell'area i, sono stati calcolati come segue:

$$RB(\tau, i) = \frac{\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_i(\tau)] - MSE[\hat{q}_i(\tau)]}{MSE[\hat{q}_i(\tau)]} \tag{4.16}$$

$$ARB(\tau, i) = \left| \frac{\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_i(\tau)] - MSE[\hat{q}_i(\tau)]}{MSE[\hat{q}_i(\tau)]} \right|, \tag{4.17}$$

dove $MSE[\hat{q}_i(\tau)]$ (4.15) è il vero valore dell'errore quadratico medio della stima del quantile di ordine τ nell'area i, $\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_i(\tau)]$ è la media delle stime dell'errore quadratico medio della stima del quantile di ordine τ nell'area i nella simulazione Monte Carlo, ovvero:

$$\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_i(\tau)] = MC^{-1} \sum_{mc=1}^{MC} \widehat{MSE}^{\cdot,\upsilon}[\hat{q}_{i,mc}(\tau)],$$

dove $\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_{i,mc}(\tau)]$ è la stima dell'errore quadratico medio (4.14) della stima del quantile di ordine τ nell'area i ottenuta nell'iterazione mc della simulazione Monte Carlo. In $\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_{i,mc}(\tau)]$ il punto, $\cdot$, indica indifferentemente la stima dell'errore quadratico medio ottenuta con l'approccio empirico o con l'approccio smooth, mentre la lettera greca υ indica che l'approccio, qualunque esso sia, è non condizionato. Il tasso effettivo di copertura è stato ottenuto calcolando la percentuale di volte in cui il vero valore è caduto all'interno dell'intervallo di confidenza ad un livello di fiducia $(1 - \alpha)$:

$$CR(q, i) = MC^{-1} \sum_{mc=1}^{MC} I(q_{i,mc}(\tau) \in CI(\tau, i, mc)), \tag{4.18}$$

dove $CI(\tau, i, mc)$ è l'intervallo di confidenza per il quantile di ordine τ nell'area i nell'iterazione mc, calcolato come $[\hat{q}_i(t) \pm z_{\alpha/2}\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_{i,mc}]^{1/2}]$, dove $z_{\alpha/2}$ è il percentile $\alpha/2$ della distribuzione normale standard e $\widehat{MSE}^{\cdot,\upsilon}[\hat{q}_{i,mc}(\tau)]$ è la stima dell'errore quadratico medio (4.14) della stima del quantile di ordine τ nell'area i ottenuta nell'iterazione mc della simulazione Monte Carlo.

Nella tabella 4.1 sono stati riportati l'errore relativo (4.16) e l'errore assoluto relativo (4.17), in percentuale, della stima dell'errore quadratico medio della stima dei quartili per piccola area per il carattere Y delle popolazioni Ω_1 e Ω_2. Vista la quantità di aree elevata (30) si è preferito riportare la

distribuzione (caratterizzata dagli indici minimi, massimo, 1^o, 2^o e 3^o quartile e media) degli indici RB (4.16) e ARB (4.17) tra le aree. Il dettaglio completo dei risultati è riportato in appendice A. Per semplificare la notazione d'ora in avanti (e nelle tabelle) si rimuove l'apice v dalla notazione $\widehat{MSE}^{\cdot,v}[\hat{q}_{i,mc}(\tau)]$ poiché si farà sempre riferimento all'approccio non condizionato, come indicato precedentemente.

Tabella 4.1: Distribuzione tra le aree dell'errore relativo (%) e dell'errore assoluto relativo (%) della stima dell'errore quadratico medio della stima dei quartili della variabile Y per le popolazioni Ω_1 e Ω_2.

		Ω_1				Ω_2			
		$\widehat{MSE}^{e}[\hat{q}_i]$		$\widehat{MSE}^{h}[\hat{q}_i]$		$\widehat{MSE}^{e}[\hat{q}_i]$		$\widehat{MSE}^{h}[\hat{q}_i]$	
		RB	ARB	RB	ARB	RB	ARB	RB	ARB
	Min.	-9.61	0.25	-12.28	0.09	-14.19	0.18	-13.51	0.30
	Qu. 1	-5.41	1.28	-4.35	1.39	-3.55	1.55	-5.45	2.27
τ	Mediana	-1.74	3.35	-2.16	2.83	0.80	3.20	-2.47	3.71
0.25	Media	-2.02	3.86	-2.44	3.82	-0.61	4.38	-2.15	4.77
	Qu. 3	0.13	6.35	-0.10	4.37	2.76	6.39	0.61	8.08
	Max.	6.41	9.61	8.32	12.28	7.93	14.19	8.84	13.51
	Min.	-11.37	0.20	-11.90	0.04	-17.60	0.57	-15.37	0.05
	Qu. 1	-6.30	2.19	-6.37	1.36	-6.66	2.80	-6.78	1.98
τ	Mediana	-3.08	4.68	-1.50	2.43	-3.51	4.51	-2.05	4.99
0.50	Media	-2.99	4.47	-2.58	3.85	-3.30	5.95	-2.22	5.16
	Qu. 3	0.01	6.30	0.98	6.37	0.80	8.63	1.68	7.20
	Max.	6.14	11.37	4.17	11.90	10.87	17.60	8.31	15.37
	Min.	-12.50	0.18	-13.34	0.04	-14.97	0.30	-16.44	0.04
	Qu. 1	-6.13	1.84	-5.79	1.32	-5.16	1.90	-5.47	3.16
τ	Mediana	-2.15	4.68	-4.59	4.59	-2.01	3.26	-1.86	4.71
0.75	Media	-2.58	4.86	-3.78	4.24	-2.02	4.56	-1.68	4.98
	Qu. 3	1.41	6.80	-0.72	5.79	1.58	6.37	3.18	6.25
	Max.	7.17	12.50	2.70	13.34	8.41	14.97	14.79	16.44

Nella tabella 4.2 si riporta l'indice di copertura CR dell'intervallo di confidenza per la stima dei quartili per piccola area con un livello di fiducia del 95% (che implica un valore di $z_{\alpha/2} = 1.96$) calcolato secondo quanto detto precedentemente. Anche in questo caso, per una maggiore chiarezza, si riporta la distribuzione (caratterizzata dagli indici minimi, massimo, 1^o, 2^o e 3^o quartile e media) dell'indice CR (4.18) tra le aree.

Dai dati delle tabelle 4.1 e 4.2 risulta evidente la buona approssimazione dello stimatore bootstrap dell'errore quadratico medio della stima dei quartili per piccola area al vero errore quadratico medio della stima dei quantili per piccola area. Anche il tasso di copertura, ottenuto utilizzando l'approssimazione normale, è molto vicino al tasso nominale fissato al 95%. I risultati sono ottimi sia per quanto riguarda l'approccio empirico sia per quanto riguarda riguarda l'approccio smooth dello schema bootstrap. Facendo un confronto tra i due approcci per la stima dell'errore del modello di superpopolazione nella procedura bootstrap, secondo la simulazione fatta, non ne emerge uno migliore tra quello empirico non condizionato e quello smooth non condizionato. Analoghi risultati (non

Tabella 4.2: Distribuzione tra le aree del tasso di copertura ottenuto dalla stima dell'errore quadratico medio della stima dei quartili della variabile Y per le popolazioni Ω_1 e Ω_2. Il livello nominale di fiducia è posto al 95%.

		Ω_1		Ω_2	
		$\widehat{MSE}^e[\hat{q}_i(\tau)]$	$\widehat{MSE}^h[\hat{q}_i(\tau)]$	$\widehat{MSE}^e[\hat{q}_i(\tau)]$	$\widehat{MSE}^h[\hat{q}_i(\tau)]$
		CR	CR	CR	CR
	Min.	0.92	0.91	0.91	0.92
	Qu. 1	0.94	0.93	0.95	0.94
τ	Mediana	0.95	0.94	0.96	0.95
0.25	Media	0.95	0.94	0.95	0.95
	Qu. 3	0.95	0.96	0.96	0.96
	Max.	0.98	0.97	0.98	0.98
	Min.	0.92	0.92	0.92	0.91
	Qu. 1	0.93	0.94	0.94	0.94
τ	Mediana	0.95	0.95	0.94	0.95
0.50	Media	0.94	0.95	0.95	0.95
	Qu. 3	0.96	0.96	0.96	0.96
	Max.	0.98	0.98	0.98	0.98
	Min.	0.92	0.92	0.90	0.90
	Qu. 1	0.94	0.93	0.94	0.94
τ	Mediana	0.94	0.95	0.94	0.96
0.75	Media	0.95	0.95	0.95	0.95
	Qu. 3	0.96	0.95	0.96	0.96
	Max.	0.97	0.98	0.98	0.98

pubblicati) sono stati ottenuti per l'approccio empirico condizionato e smooth condizionato. I risultati ottenuti erano senz'altro attesi se si considera la popolazione Ω_1, costruita con disturbi simmetrici (normali), mentre non erano scontati per la popolazione Ω_2, generata con disturbi asimmetrici (χ^2_3). Inoltre, si poteva ipotizzare una migliore performance dell'approccio bootstrap di tipo smooth (condizionato o non condizionato) che invece non è risultato preponderante, almeno secondo gli indici rilevati, sullo stimatore bootstrap basato sull'approccio empirico. Tuttavia si sottolinea che, al contrario dell'approccio empirico, nell'approccio smooth, sia esso condizionato o non condizionato, si rispettano i teoremi (proposti da Lombardia *e altri* (2003)) su cui si basa lo stimatore bootstrap proposto. E' evidente, comunque, come nell'ambito della stima per piccole aree non si possa dare troppo peso ai comportamenti asintotici degli stimatori, dunque, vista anche la maggior velocità di calcolo e i buoni risultati ottenuti, si consiglia di non scartare l'approccio empirico dello schema bootstrap presentato.

4.5 Simulazione Design-based per lo stimatore bootstrap dell'errore quadratico medio dello stimatore CD per la funzione di ripartizione per piccola area

Per la simulazione basata su disegno è stato utilizzato il campione Leaving Standard Measurement Survey Albania del 2003 (LSMS Albania), d'ora in avanti campione Albania. Una volta verificato il comportamento dello stimatore bootstrap proposto per dati artificiali, totalmente controllati a priori, si vuole analizzare in questo paragrafo il comportamento dello stimatore bootstrap dell'errore quadratico medio della stima dei quantili per piccola area con dati reali. Solitamente una simulazione basata su disegno utilizza una popolazione sintetica generata da un campione di dati reali. Ciò sarebbe stato assolutamente possibile anche con il campione Albania del 2003. Purtroppo, a causa della lentezza computazionale della procedura di stima bootstrap che è stata proposta e delle dimensioni della ipotetica popolazione sintetica, non è stato possibile agire nel modo consueto; per fare una simulazione Monte Carlo di 200 iterazioni, con i parametri bootstrap B ed L che sono poi stati effettivamente usati, sarebbe stato necessario un tempo troppo lungo[10]. Visto che non è possibile agire sui parametri della procedura bootstrap (già ridotti al minimo secondo l'opinione dell'autore) è stato deciso di trattare il campione Albania come se fosse una popolazione. In questo modo è stato possibile effettuare le simulazioni Monte Carlo in tempi accettabili. La variabile di interesse, Y, è il *reddito equivalente* e si considera nota solo a livello campionario[11]. Le variabili ausiliarie che sono state utilizzate sono *household diploma*, la frazione di diplomati sul totale dei membri della famiglia, *household male*, la frazione dei maschi sul totale dei membri della famiglia, e *household land*, variabile *dummy* sul possesso di un terreno[12]. Le variabili ausiliarie si considerano note in modo certo e senza errore per tutte le unità della popolazione. Il campione Albania, la nostra popolazione, è composto da 3591 unità divise in 36 aree. In ogni area è stato estratto un campione casuale semplice di dimensione pari a 1/10 della dimensione della popolazione dell'area stessa; nelle aree in cui il numero di unità è risultato inferiore a 50 è stato estratto un campione di 5 unità. Il campione estratto è risultato di 398 unità. Nelle simulazioni Monte Carlo la popolazione è fissa mentre un campione è estratto ad ogni iterazione. Su ogni campione estratto si è stimato il quartile di piccola area per la variabile di interesse e il relativo errore quadratico medio. Il quartile per piccola area è stato stimato con lo stimatore CD utilizzando come modello di superpopolazione il modello di regressione M-quantile. L'errore quadratico medio della stima del quartile per piccola area è stato stimato utilizzando la procedura bootstrap con approccio smooth non condizionato. Con l'approccio empirico non condizionato sono stati ottenuti risultati simili (non pubblicati). L'approccio condizionato, sia esso smooth o empirico, ha prodotto dei risultati, in termini di errore relativo, errore assoluto relativo e livello di copertura, accettabili in alcune aree e del tutto inaccettabili in altre aree. D'altronde si deve considerare che in 22 aree la numerosità campionaria è minore di 10 unità: un numero di osservazioni troppo limitato per ottenere delle stime utilizzabili dell'errore del modello di superpopolazione, necessarie per l'approccio condizionato, sia esso smooth o empirico.

La simulazione Monte Carlo è stata fatta con 200 iterazioni, mentre per quanto riguarda i parametri dello stimatore bootstrap è stato posto $L = 500$ e $B = 1$, dove L indica il numero di campioni estratti dalla popolazione artificiale bootstrap, $\Omega^{*b}_{Alb}, b = (1, \ldots, B)$, e B indica il numero delle popolazioni bootstrap generate. Purtroppo non è stato possibile utilizzare valori maggiori per il parametro B dello

[10] Si tratta di circa 30-50 giorni per ogni approccio.

[11] Per un approfondimento sui dati del campione LSMS Albania si consulti Grosh e Glewwe (1995).

[12] Inglesismo per definire una variabile indicatrice di tipo 0,1. 0 la famiglia non possiede il terreno, 1 la famiglia possiede il terreno.

stimatore bootstrap e per il numero delle simulazioni Monte Carlo per motivi di tempo; sarebbe stata auspicabile una simulazione Monte Carlo con 500 (se non 1000) iterazioni e un valori di B pari a 50.

Per la stima dell'errore quadratico medio della stima dei quartili per piccola area, ottenuta con l'approccio smooth non condizionato, è stato calcolato l'errore relativo (4.16), l'errore assoluto relativo (4.17) e il tasso di copertura (4.18) seguendo le indicazioni specificate nel paragrafo precedente. Gli indici specificati sono stati riportati nella tabella 4.3. Vista la quantità di aree elevata (36), si è preferito riportare la distribuzione degli indici RB(4.16), ARB (4.17) e CR (4.18) tra le aree. I risultati dettagliati sono stati riportati in appendice A

Tabella 4.3: Distribuzione tra le aree dell'errore relativo (%) e dell'errore assoluto relativo (%) della stima dell'errore quadratico medio della stima dei quartili del reddito disponibile per la popolazione Albania e del tasso di copertura ottenuto dalla stima dell'errore quadratico medio tramite approssimazione normale con una fiducia nominale del 95%. Approccio smooth non condizionato.

		RB	ARB	CR
	Min.	-47.05	1.08	0.58
	Qu. 1	-17.16	7.35	0.90
τ	Mediana	-1.53	16.62	0.95
0.25	Media	4.43	25.56	0.92
	Qu. 3	11.35	31.99	0.97
	Max.	108.80	108.80	1.00
	Min.	-46.11	0.27	0.70
	Qu. 1	-19.04	8.73	0.90
τ	Mediana	-4.35	17.95	0.95
0.50	Media	10.80	31.08	0.93
	Qu. 3	11.46	28.92	0.97
	Max.	290.80	290.80	1.00
	Min.	-30.94	0.08	0.84
	Qu. 1	-12.68	11.00	0.93
τ	Mediana	4.91	17.94	0.97
0.75	Media	11.84	24.22	0.95
	Qu. 3	23.27	26.94	0.98
	Max.	189.60	189.60	1.00

Osservando i dati raccolti nella tabella 4.3 si nota che l'errore relativo (RB) vale in media tra il 4% e l'11%, mentre la mediana varia tra -4% e 5%, discostandosi in maniera sensibile dalla media. Infatti dai valori minimo e massimo della distribuzione dell'errore relativo (tra le 36 aree della popolazione di riferimento) si rilevano valori anomali, segno che in certe aree la stima dell'errore quadratico medio non ha approssimato in modo corretto il vero errore quadratico medio[13]. Anche guardando ai valori dell'errore assoluto relativo (ARB) si confermano i risultati appena esposti. Inoltre media e mediana hanno valori sensibilmente più alti rispetto all'errore relativo, ciò è positivo perché indica che la distorsione dello stimatore dell'errore quadratico medio tra le aree si compensa. L'errore assoluto relativo medio più alto è di poco superiore al 30% (31.08%), ciò indica che, nonostante in alcune aree non si siano ottenuti buoni risultati, complessivamente il metodo di stima bootstrap proposto ha

[13]Si ricorda che per "vero" errore quadratico medio di intende quello ottenuto dalla simulazione Monte Carlo, come specificato nella (4.15).

prodotto risultati accettabili. Se si considera la ridottissima numerosità campionaria in molte aree (in 22 aree il campione è minore di 10 unità e, tra queste, in 18 aree è minore di 5 unità) allora possiamo considerare i risultati ottenuti molto soddisfacenti. Il tasso di copertura è ottimo e si attesta in media e in mediana molto vicino al valore nominale di 0.95.

Le aree in cui è stata riscontrata, in particolare, una pessima approssimazione del vero errore quadratico medio sono la 2, la 13 e la 29. In queste aree il modello si superpopolazione utilizzato non riesce a prevedere correttamente per alcune unità la variabile di studio. Ciò comporta che per quelle unità in cui il valore della Y non è stato predetto correttamente il residuo assuma valori sballati rispetto agli altri residui. Nella fase di campionamento dei residui, nell'approccio non condizionato, non viene associato a quelle unità particolari un residuo adeguato, per questo motivo si presume che la procedura bootstrap non funzioni come dovrebbe. I problemi potrebbero essere anche altri, legati alla stima del quantile medio di area, tuttavia con gli strumenti diagnostici attualmente disponibili per il modello di regressione M-quantile è difficile indagare sui motivi del comportamento anomalo rilevato in queste tre aree in particolare. Se si escludono le aree 2, 13 e 29 i risultati della simulazione in termini di errore relativo ed errore assoluto relativo sono nettamente migliori e in linea con le aspettative. Nella tabella 4.4 sono riportati l'errore relativo ,RB (4.16), l'errore assoluto relativo, ARB (4.17) e il tasso di copertura, CR (4.18) togliendo dai risultati le aree 2, 13 e 29. I risultati sono riportati come distribuzione degli indici RB(4.16), ARB (4.17) e CR (4.18) tra le (33) aree della popolazione di riferimento.

Tabella 4.4: Distribuzione tra le aree dell'errore relativo (%) e dell'errore assoluto relativo (%) della stima dell'errore quadratico medio della stima dei quartili del reddito disponibile per la popolazione Albania e del tasso di copertura ottenuto dalla stima dell'errore quadratico medio tramite approssimazione normale con una fiducia nominale del 95%. Approccio smooth non condizionato.

		RB	ARB	CR
	Min.	-47.05	1.08	0.58
	Qu. 1	-18.39	6.14	0.90
τ	Mediana	-5.03	15.85	0.95
0.25	Media	-3.27	19.78	0.92
	Qu. 3	6.14	28.85	0.97
	Max.	89.04	89.04	1.00
	Min.	-46.11	0.27	0.70
	Qu. 1	-19.28	8.31	0.90
τ	Mediana	-7.52	13.31	0.95
0.50	Media	-3.99	18.13	0.93
	Qu. 3	8.31	22.58	0.97
	Max.	62.36	62.36	1.00
	Min.	-30.94	0.08	0.84
	Qu. 1	-13.45	12.00	0.93
τ	Mediana	4.07	17.69	0.97
0.75	Media	6.24	19.39	0.95
	Qu. 3	20.47	26.33	0.98
	Max.	70.76	70.76	1.00

Dai valori di RB e ARB riportati nella tabella 4.4 si può constatare un sostanziale miglioramento

rispetto ai dati della tabella 4.3. La media dell'errore relativo RB tra le 33 aree della popolazione Albania è -3.27%, -3.99% e 6.25% rispettivamente per il 1^o, 2^o e 3^o quartile, mentre l'errore assoluto relativo medio per i quartili è rispettivamente 19.78%, 18.13% e 19.39%, dunque sempre inferiore al 20%. E' vero, come si può notare dai valori agli estremi della distribuzione di RB e ARB, che ci sono ancora delle aree in cui l'approssimazione bootstrap del vero errore quadratico medio non è accettabile ma in complesso i risultati sono soddisfacenti. Anche esaminando i valori della mediana della distribuzione tra le aree dell'errore relativo (-5.03%, -7.52% e 4.07%, rispettivamente per il 1^o, 2^o e 3^o quartile) e dell'errore assoluto relativo (15.85%, 13.31% e 17.69%, rispettivamente per il 1^o, 2^o e 3^o quartile) si conclude che lo stimatore bootstrap proposto approssimi (abbastanza) correttamente il vero errore quadratico medio della stima dei quartili per piccola area.

4.6 Applicazione dello stimatore bootstrap dell'errore quadratico medio dello stimatore CD per la funzione di ripartizione per piccola area alla stima per provincia dei quartili del reddito equivalente disponibile in toscana

Le indagini su vasta scala, come l'indagine EU-SILC (acronimo di *European Union - Statistics on Income and Living Conditions*), sono progettate per fornire stime affidabili per un dato livello territoriale o per certi domini. L'indagine EU-SILC in Italia è progettata per ottenere stime attendibili a livello regionale. Vista la volontà di enti pubblici e privati di avere stime a livelli territoriali sempre più dettagliati ci si è posti l'obiettivo di ottenere delle stime attendibili a livello provinciale. La variabile che si intende esaminare è il reddito equivalente disponibile, rilevato tramite l'indagine EU-SILC del 2004 (con riferimento ai redditi del 2003). L'output di questa applicazione sarà la stima dei quartili, e del loro errore standard, del reddito equivalente disponibile per provincia nella regione Toscana nell'anno 2003. Poiché con le stime dirette per provincia, che è un'unità territoriale non prevista nel disegno di campionamento, non si ottengono risultati sufficientemente affidabili il problema rientra nell'ambito della stima per piccole aree. Dato che tutti i modelli di stima per i quantili per piccola area che sono stati esposti in questo lavoro richiedono l'uso di variabili ausiliarie note in modo certo e senza errore per tutte le unità della popolazione, si è scelto di utilizzare il Censimento *Famiglie e Abitazioni* del 2001 da cui si sono state scelte le variabili ausiliarie che sono state ritenute necessarie.

L'indagine EU-SILC nasce per armonizzare le statistiche in ambito socio-economico all'interno dell'Unione Europea. Il Regolamento del Parlamento europeo, Statistics on Income and Living Conditions (n. $1177/2003$, meglio noto come Eu-Silc), risponde alla crescente domanda di informazioni da parte delle istituzioni nazionali ed europee, della comunità scientifica e degli stessi cittadini sulle condizioni di vita nei diversi paesi dell'Unione. Il progetto ha come obiettivo principale la produzione sistematica di statistiche comunitarie su reddito, povertà ed esclusione sociale, sia a livello trasversale che longitudinale, puntando all'armonizzazione di un insieme di indicatori statistici (Ceccarelli *e altri*, 2008).

Gli aspetti metodologici del nuovo strumento Eu-Silc sono stati sviluppati in cinque regolamenti della Commissione (Sampling and tracing rules; Definitions, List of primary variables; Fieldwork aspect and imputation procedures; Intermediate and final quality reports); inoltre, ogni anno, un nuovo regolamento definisce la lista di variabili target per un modulo ad hoc scelto tra una rosa di tematiche di interesse. Il progetto ha previsto un periodo di transizione (fino al 2007) che consentisse agli Istituti nazionali di statistica di adattare i propri strumenti agli standard comuni con particolare riferimento a: i fitti imputati, i contributi sociali a carico dei datori di lavoro e le componenti del reddito lordo.

Il progetto Eu-Silc è stato lanciato nel 2003 su base sperimentale in sette paesi (Belgio, Norvegia, Grecia, Lussemburgo, Austria, Danimarca e Irlanda) anche con l'obiettivo di studiare l'impatto dei cambiamenti dello strumento nella serie degli indicatori di Laeken. Il lancio ufficiale di Eu-Silc si è avuto invece nel 2004 in tredici dei vecchi Stati membri, compresa l'Italia (non hanno partecipato Paesi Bassi, Germania e Regno Unito), e in dieci di quelli nuovi (eccetto Estonia), oltre che in Norvegia e Islanda. Nel 2005, Eu-Silc ha raggiunto la sua piena estensione con venticinque stati membri, più Norvegia e Islanda (Eu-Silc è in preparazione anche in Turchia, Romania, Bulgaria, Svizzera) (Ceccarelli *e altri*, 2008).

Nell'indagine EU-SILC italiana, la popolazione di riferimento è costituita da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente all'estero. Sono escluse le famiglie residenti in Italia che vivono abitualmente all'estero e i membri permanenti delle convivenze istituzionali (ospizi, brefotrofi, istituti religiosi, caserme, eccetera).

L'unità di rilevazione è la famiglia di fatto. Questa va intesa come un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti ed aventi dimora abituale nello stesso comune (anche se non residenti secondo l'anagrafe nello stesso domicilio).

Il disegno di campionamento adottato dall'ISTAT per l'indagine EU-SILC segue uno schema standard a due stadi comuni-famiglie con stratificazione dei comuni in base alla dimensione demografica; lo schema, ormai consolidato, viene utilizzato per le principali indagini ISTAT sulle famiglie condotte mediante intervista faccia a faccia. Il disegno di campionamento adottato nell'ambito delle regioni prevede un disegno a due stadi comuni-famiglie in cui i comuni vengono suddivisi in "auto-rappresentativi" e "non auto-rappresentativi"; inoltre i comuni non auto-rappresentativi sono stratificati in base alla dimensione demografica.

Nell'ambito dell'insieme dei comuni auto-rappresentativi, ciascun comune viene considerato come uno strato a se stante e viene adottato un disegno noto con il nome di campionamento a grappoli. Le unità primarie di campionamento sono rappresentate dalle famiglie anagrafiche, estratte in modo sistematico dall'anagrafe del comune stesso; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto. Nell'ambito dei comuni non auto-rappresentativi viene adottato un disegno a due stadi con stratificazione delle unità primarie. Le unità primarie sono i comuni, le unità secondarie sono le famiglie anagrafiche, estratte sistematicamente come per i comuni auto-rappresentativi; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto. I comuni vengono selezionati con probabilità proporzionali alla loro dimensione demografica e senza reimmissione, mentre le famiglie vengono estratte con probabilità uguali e senza reimmissione.

Nell'indagine in esame, i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- auto-ponderazione del campione a livello regionale;
- selezione di quattro comuni campione nell'ambito di ciascuno strato definito sui comuni dell'insieme non auto-rappresentativi;
- scelta di un numero minimo di famiglie da intervistare in ciascun comune campione; tale numero è stato posto pari a ventiquattro;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Il procedimento di stratificazione, attuato all'interno di ogni regione geografica, si articola nelle seguenti fasi:

- ordinamento dei comuni del dominio in ordine decrescente secondo la loro dimensione demografica in termini di popolazione residente;
- determinazione di una soglia di popolazione per la definizione dei comuni auto-rappresentativi, mediante la relazione:

$$\lambda_i = \frac{\bar{m}_i \delta_i}{f_i}$$

  dove per la generica regione i si è indicato con $\bar{m}_i$ il numero minimo di famiglie da intervistare in ciascun comune campione, con δ_i il numero medio dei componenti per famiglia e con f_i la frazione di campionamento;
- suddivisione di tutti i comuni nei due sottoinsiemi auto-rappresentativi e non auto-rappresentativi: i comuni di dimensione superiore o uguale a λ_i sono definiti come comuni auto-rappresentativi e i rimanenti come non auto-rappresentativi;
- suddivisione dei comuni dell'insieme non auto-rappresentativi in strati aventi dimensione, in termini di popolazione residente, approssimativamente costante e all'incirca pari a 4xλ_i.

Effettuata la stratificazione, i comuni auto-rappresentativi sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni non auto-rappresentativi, nell'ambito di ogni strato vengono estratti quattro comuni campione con probabilità proporzionali alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow (1949). La selezione delle famiglie da intervistare in ogni comune campione viene effettuata dalla lista anagrafica di ciascun comune senza reimmissione e con probabilità uguali.

La numerosità del campione EU-SILC nazionale è stata fissata sulla base della precisione delle stime della percentuale di famiglie povere nell'ipotesi di campionamento casuale semplice. La numerosità campionaria minima, definita da EUROSTAT, per l'Italia p risultata pari a 7250 famiglie. Ogni paese dell'Unione ha dovuto stabilire la numerosità campionaria effettivamente da selezionare sulla base della valutazione dell'effetto del disegno relativo al disegno di campionamento prescelto, tenendo in considerazione l'impatto sull'efficienza delle stime prodotto dalla stratificazione, dal clustering e dalla ponderazione, e dei tassi attesi di mancata risposta totale. La numerosità a livello di ripartizione, definita tenendo conto dell'effetto del disegno, è stata allocata tra le regioni seguendo un'ottica di compromesso tra un'allocazione uniforme e un'allocazione proporzionale alla popolazione delle regioni, tale da garantire sia l'affidabilità prefissata delle stime a livello nazionale e ripartizionale, sia quella delle stime a livello regionale (Ceccarelli *e altri*, 2008).

Dunque non è stato tenuto conto della ripartizione provinciale nel disegno di campionamento. La numerosità campionaria nella regione Toscana è di 1751 famiglie, ed è ripartita tra le province da un minimo di 70 famiglie (provincia di Grosseto) ad un massimo di 545 famiglie (provincia di Firenze). I dati in dettaglio sula numerosità campionaria sono riportati nell'appendice A, tebella A.17. Si ritiene necessario, visto quanto detto a proposito del campione EU-SILC, utilizzare i metodi di stima per piccola area al fine di ottenere una stima dei quartili del reddito equivalente disponibile per provincia.

La rilevazione del reddito a livello familiare è, al contrario di quanto si possa credere, un fenomeno molto complesso da rilevare. Per migliorare le tecniche di rilevazione del reddito in relazione ai possibili problemi psicologici, di memoria, di conoscenza e di motivazione dei rispondenti, è stata

condotta, infatti, una serie di colloqui in profondità, con l'obiettivo di far emergere le propensioni, i condizionamenti, le resistenze comuni alle differenti tipologie di percettori. Ai colloqui in profondità è seguito uno studio - condotto con la tecnica del focus group e specificamente rivolto ai lavoratori autonomi - allo scopo di individuare, attraverso la discussione e le valutazioni emerse dai partecipanti, le strategie di approccio e di raccolta dei dati (contatto con la famiglia, tipo e sequenza delle domande, eccetera) più adeguate ed efficaci a massimizzare la disponibilità dei rispondenti e l'attendibilità delle loro risposte (Ceccarelli *e altri*, 2008).

Il reddito è definito, secondo gli economisti, da $Y = C + S$: il reddito è uguale ai consumi più il risparmio. Nell'indagine Eu-Silc, il reddito viene osservato come un insieme di entrate ricavate da fonti diverse, secondo lo schema seguente (Ceccarelli *e altri*, 2008):

- Reddito guadagnato sul mercato
  - Redditi da lavoro
    - dipendente
    - autonomo
  - Redditi da capitale
    - reale (affitti e rendite di terreni e fabbricati)
    - finanziario (interessi, dividendi, utili)
    - intellettuale (diritti d'autore)
- Reddito da trasferimenti
  - Trasferimenti pubblici
    - pensioni
    - altri trasferimenti pubblici in denaro (per esempio assegni familiari)
  - Trasferimenti privati
    - aiuti in denaro di familiari ed amici, assegni di ex-coniugi
    - aiuti in denaro di istituzioni private (per esempio da associazioni religiose)

Accanto a queste componenti, misurate in moneta, si considerano anche altre risorse "non-monetarie" che concorrono al benessere familiare:

- salari in natura (fringe benefits): come l'uso gratuito di una abitazione, l'auto aziendale per usi privati, i buoni pasto (dal 2007), l'asilo nido aziendale (dal 2007);
- affitti imputati delle case occupate dai proprietari: che è pari al valore del servizio che queste abitazioni rendono a chi ne è proprietario. Per convenzione, è "come se" i proprietari affittassero la casa a sé stessi
- autoconsumi (dal 2007): cioè il valore stimato dei beni che la famiglia ha eventualmente prodotto per il proprio consumo, come per esempio frutta, vino e ortaggi.

Sono invece escluse dalla definizione di reddito adottata per l'indagine Eu-Silc, per difficoltà di rilevazione e/o di stima del valore monetario corrispondente, alcune componenti che pure concorrono a determinare le condizioni economiche delle famiglie:

- trasferimenti pubblici in natura, come per esempio i servizi sanitari e scolastici forniti gratuitamente o a prezzi agevolati dalla pubblica amministrazione. È indiscutibile che tali beni, se utilizzati, contribuiscono al benessere delle famiglie. Tuttavia, è praticamente impossibile, in assenza di un mercato privato indipendente, stimare il valore corrispondente a questo tipo di benefici pubblici ricevuti in natura dalle famiglie. Anche quando esistono servizi privati in concorrenza di quelli pubblici, infatti, il loro prezzo di mercato è "residuale" rispetto alla politica di offerta dell'operatore pubblico. La valutazione al costo di produzione, a sua volta, ignora la qualità dei servizi e può non riflettere la disponibilità a pagare degli utenti
- i beni e i servizi in natura ricevuti da parenti e amici (per esempio, la cura dei figli da parte di una parente non coabitante), per la difficoltà di valutarne sia la quantità, sia il valore "figurativo"
- per difficoltà di rilevazione, sono anche escluse tutte quelle attività lavorative effettuate dai membri della famiglia in sostituzione di analoghi servizi di mercato, come per esempio la riparazione di elettrodomestici, la manutenzione di mobili eccetera. La difficoltà in questo caso riguarda la vasta gamma coperta dalla produzione domestica: in effetti, anche le pulizie di casa e la preparazione dei cibi sostituiscono servizi acquistabili altrimenti sul mercato

Sono, infine, escluse alcune entrate eccezionali che sono considerate come variazioni "istantanee" della ricchezza:

- le vincite alla lotteria
- le eredità e le donazioni una tantum
- i guadagni in conto capitale, cioè gli aumenti del valore del patrimonio posseduto (case, terreni, gioielli, azioni ed altre attività finanziarie)

Oltre alla complessità della definizione di reddito si aggiungono altri tipi di problemi all'indagine EU-SILC. Infatti la rilevazione campionaria dei redditi pone numerosi problemi, dovuti a due ordini di motivi:

- scarsa conoscenza da parte degli intervistati:
  - delle definizioni di reddito
  - degli importi esatti percepiti
- scarsa disponibilità a rispondere all'intervista:
  - per diffidenza (soprattutto timore di controlli fiscali)
  - per sfiducia nelle istituzioni e nell?utilità delle indagini statistiche

In Italia l'ISTAT, che gestisce l'indagine EU-SILC, ha approntato strategie ad hoc per affrontare i problemi sopracitati. La descrizione dell'indagine che è stata fatta non è e non vuole essere esaustiva, l'indagine EU-SILC ha una struttura molto complessa. Per un approfondimento sull'indagine EU-SILC in Italia si consiglia di consultare Ceccarelli *e altri* (2008).

La variabile di interesse oggetto di studio in questo lavoro è il reddito equivalente disponibile. Il reddito equivalente disponibile, nell'indagine EU-SILC, è così definito:

$$\begin{aligned}&\text{Reddito Equivalente Disponibile} = \\ &= \frac{\text{Reddito Familiare Disponibile Totale} \cdot \text{Fattore di Inflazione per Non-risposta}}{\text{n. Membri Equivalenti}},\end{aligned} \tag{4.19}$$

dove il Reddito Familiare Disponibile è uguale al reddito lordo al netto di contributi previdenziali e assistenziali, reddito dei minori di anni 16, tasse sul welfare, trasferimenti all'interno del nucleo familiare e delle imposte sui redditi. Il reddito lordo è dato dalla somma delle voci elencate a pagina 85. Il Fattore di Inflazione per Non-risposta equivale al reddito imputato in caso di non risposta di uno o più membri all'interno della famiglia intervistata. Per un approfondimento sui meccanismi di imputazione per questa variabile si consulti la documentazione ufficiale dell'indagine EU-SILC. Il n. Membri Equivalenti indica il numero dei componenti del nucleo familiare opportunamente pesati per considerare le economie di scala familiari. Sia $M14_+$ il numero dei membri della famiglia che hanno compiuto 14 anni nell'anno di riferimento dell'indagine (nel caso in esame il 2003) e sia $M13_-$ il numero dei membri della famiglia che hanno 13 anni o meno nell'anno di riferimento dell'indagine (2003 nel caso in esame), allora il n. Membri Equivalenti $= 1 + 0.5 \cdot (M14_+ - 1) + 0.3 \cdot (M13_-)$.

Per la variabile ausiliaria, che deve essere nota in modo certo e senza errore per tutte le unità della popolazione, sono stati utilizzati i dati censuari. Il censimento disponibile più recente sulle famiglie della regione Toscana è il censimento *Famiglie e Abitazioni* del 2001, curato dall'ISTAT[14]. Il censimento garantisce un grado di dettaglio territoriale (fino al comune e alla sezione di censimento) non deducibile da nessun'altra fonte statistica o amministrativa. I principali obiettivi del Censimento sono (fonte ISTAT):

- Il conteggio della popolazione e la rilevazione delle sue caratteristiche strutturali
- L'aggiornamento e la revisione delle anagrafi
- La determinazione della popolazione legale
- La raccolta di informazioni sulla consistenza numerica e sulle caratteristiche strutturali delle abitazioni

In un contesto ideale, per prevedere la variabile di interesse reddito equivalente disponibile, sarebbe auspicabile utilizzare come covariate l'età (o fasce di età), l'età al quadrato, il sesso, il titolo di studio (o gli anni di studio), il gruppo etnico[15], il numero di componenti della famiglia e la zona geografica di residenza della famiglia, che dovrebbero essere rilevate per tutte le unità della popolazione di riferimento nell'anno 2004. Purtroppo, non è stato possibile utilizzare un modello del genere. I motivi sono principalmente due: **i.** l'unità di rilevazione dell'indagine EU-SILC è la famiglia e dunque non è univoca la scelta delle variabili età, sesso, titolo di studio e gruppo etnico, in riferimento ad una famiglia e **ii.** le covariate ottenute dall'ultimo censimento famiglie e abitazioni sono riferite al 2001 e non al 2004. In risposta al problema **i.** è stato deciso di utilizzare come covariata solo il numero dei componenti della famiglia, includendo nel modello anche una costante. Come zona geografica di residenza della famiglia è stata utilizzata la Provincia, la piccola area oggetto di studio, che è stata inclusa nel processo di stima per determinare gli effetti di area. Per il problema dello sfasamento temporale tra la variabile ausiliaria, numero dei componenti della famiglia, nel Censimento 2001 e nell'indagine EU-SILC del 2004 si è ipotizzato che non ci fosse una differenza rilevante. Questa assunzione è parzialmente confermata da un test di verifica di ipotesi per la differenza tra medie effettuato tra la media del numero dei componenti della famiglia nel Censimento 2001 (dato certo) e la media del numero dei componenti della famiglia nell'indagine EU-SILC del 2004 per le province della regione Toscana. Tale differenza è risultata significativa solo per le province di Massa-Carrare e Arezzo ad un livello di

[14] Istituto Nazionale di Statistica.

[15] forse poco rilevante in Italia ma sicuramente da considerare in futuro.

fiducia del 95% e per nessuna provincia ad un livello di fiducia del 90%. Il modello di superpopolazione che è stato utilizzato per la stima dei quartili del reddito equivalente disponibile nell'anno 2003, definito secondo la (4.19), per le province della regione Toscana è il seguente modello di regressione M-quantile:

$$y_{ji} = \alpha_\psi(\theta_i) + \beta_\psi(\theta_i)x_{ji} + \varepsilon_{ji},$$

dove y_{ji} è il reddito equivalente disponibile (4.19) della j-esima famiglia della provincia i, con $i = (1, \ldots, 10)$, x_{ji} è il numero dei componenti della j-esima famiglia della provincia i, ε_{ji} è un errore indipendente e identicamente distribuito per la famiglia j dell'area i, $\alpha_\psi(\theta_i)$ e $\beta_\psi(\theta)_i)$ sono i coefficienti di regressione (incogniti), ψ è la funzione di perdita Huber 2 descritta nel capitolo 2 e $\hat{\theta}_i$ è il coefficiente M-quantile della provincia i (capitolo 2).

Utilizzando lo stimatore della funzione di ripartizione per piccola area (3.27) si è ottenuta la stima dei quartili (3.36) del reddito equivalente disponibile del 2003 per le province della regione Toscana, denominata d'ora in avanti $\hat{q}_i(\tau)$, con $\tau = (0.25, 0.50, 0.75)$. Avvalendosi della procedura bootstrap descritta nel paragrafo 4.3 si è ottenuta la stima dell'errore quadratico medio della stima $\hat{q}_i(\tau)$. L'approccio bootstrap utilizzato è stato quello empirico non condizionato. La scelta è caduta su tale schema per due motivi: **i.** l'approccio empirico ha dato nelle simulazioni Monte Carlo risultati molto simili all'approccio smooth richiedendo meno tempo macchina e **ii.** l'approccio non condizionato è risultato preferibile, in base alle considerazioni fatte nel paragrafo 4.4 e 4.5, rispetto all'approccio condizionato[16]. Nello schema bootstrap utilizzato per la stima dell'errore quadratico medio di $\hat{q}_i(\tau)$ è stata generata una sola popolazione bootstrap Ω^* ($B = 1$) dalla quale sono stati estratti 500 campioni bootstrap s^* ($L = 500$).

I risultati ottenuti sono riportati nella tabella 4.5. Per ogni provincia è stata riportata la stima (ottenuta secondo la (3.27) e la (3.36)) dei quartili del reddito equivalente disponibile $\hat{q}_i(\tau), \tau = (0.25, 0.50, 0.75)$ e la stima dell'errore standard di $\hat{q}_i(\tau), \tau = (0.25, 0.50, 0.75)$, denominata Std. Err $(\hat{q}_i(\tau))$.

Tabella 4.5: Stima dei quartili, e del relativo errore standard, del reddito equivalente disponibile nel 2003 per provincia (Toscana).

Provincia	$\hat{q}_i(0.25)$	Std.Err($\hat{q}_i(0.25)$)	$\hat{q}_i(0.50)$	Std.Err($\hat{q}_i(0.50)$)	$\hat{q}_i(0.75)$	Std.Err($\hat{q}_i(0.75)$)
Massa	9411	598	14393	714	19126	1127
Lucca	10668	623	14876	759	20309	1238
Pistoia	10088	555	15163	693	20797	1096
Firenze	11486	306	15588	363	21865	596
Livorno	11275	564	15105	646	21310	1176
Pisa	11100	549	14950	656	20758	1102
Arezzo	12310	525	15735	601	20862	1055
Siena	11058	553	15312	680	21416	1164
Grosseto	11592	826	15420	953	21322	1632
Prato	10886	558	15319	612	21077	1074

[16]Vista la dimensione del data set utilizzato e la potenza di calcolo disponibile si è preferito scegliere l'approccio empirico rispetto a quello smooth che richiede molto tempo per la stima del parametro di bandwidth h.

Nella tabella 4.6 sono stati riportati i quartili campionari per provincia del campione EU-SILC, da usare come termine di paragone per le stime per piccole aree ottenute.

Tabella 4.6: Quartili campionari (EU-SILC 2004, redditi 2003) del reddito equivalente disponibile per provincia (Toscana).

Provincia	Qu. 1	Mediana	Qu. 3
Massa	9846	13713	18896
Lucca	9920	14982	20573
Pistoia	10143	15182	20945
Firenze	11677	15822	21865
Livorno	11207	15198	21118
Pisa	11191	14744	20794
Arezzo	12331	15731	20998
Siena	11168	15128	22150
Grosseto	11574	15198	21114
Prato	10615	15823	21235

Nella tabella 4.7 è stato riportato il coefficiente di variazione per provincia dello stimatore per piccola area ($\hat{q}_i(\tau)$), $CV = \hat{q}_i(\tau)/\text{Std. Err. } (\hat{q}_i(\tau)), \tau = (0.25, 0.50, 0.75)$.

Tabella 4.7: Coefficiente di variazione dello stimatore $\hat{q}_i(\tau)$ per provincia (Toscana).

Provincia	CV($\hat{q}_i(0.25)$)	CV($\hat{q}_i(0.50)$)	CV($\hat{q}_i(0.75)$)
Massa-Carrara	0.064	0.050	0.059
Lucca	0.058	0.051	0.061
Pistoia	0.055	0.046	0.053
Firenze	0.027	0.023	0.027
Livorno	0.050	0.043	0.055
Pisa	0.049	0.044	0.053
Arezzo	0.043	0.038	0.051
Siena	0.050	0.044	0.054
Grosseto	0.071	0.062	0.077
Prato	0.051	0.040	0.051

Le stime dei quartili del reddito equivalente disponibile presentate nella tabella 4.5 sono in linea con i quartili per provincia del campione EU-SILC (riportati nella tabella 4.6), che, anche essendo stime inconsistenti, rappresentano un ottimo termine di paragone per le stime per piccola area ottenute con l'utilizzo di modelli di superpopolazione. La stima dell'errore standard dello stimatore $\hat{q}_i(\tau)$ ha valori che crescono dal primo al terzo quartile in tutte le province, ma analizzando il coefficiente di variazione (tabella 4.7) si osserva una certa stabilità tra i quartili delle province. Per la provincia di Firenze l'errore standard è sensibilmente più basso rispetto a quello delle altre province; tuttavia questo è un risultato atteso vista la numerosità nel campione EU-SILC molto più elevata nella provincia di Firenze rispetto alle altre province. Per una migliore rappresentazione del fenomeno studiato si è ritenuto utile riportare la stima dei quartili del reddito equivalente disponibile per provincia su una carta geografica della toscana (figure 4.1, 4.2 e 4.3).

In conclusione le stime dei quartili del reddito disponibile per provincia in Toscana per l'anno 2003 e del relativo errore standard risultano stabili, nel senso che non ci sono valori sensibilmente diversi dalle stime campionarie e con un errore standard contenuto, nel senso che si reputa che le stime proposte siano affidabili.

Si ringrazia il Dipartimento di Metodi Quantitativi della Facoltà di Economia, Università di Siena, per aver gentilmente concesso i dati del censimento Famiglie e Abitazioni del 2001 e i dati dell'indagine EU-SILC del 2004. I dati sono stati forniti solo per le variabili che sono state menzionate in questa tesi e unicamente per essere usati in questa tesi e non in altri lavori, di qualunque genere essi siano.

Figura 4.1: Stima dei quartili, e del relativo errore standard, del reddito equivalente disponibile nel 2003 per provincia (Toscana).

Figura 4.2: Stima dei quartili, e del relativo errore standard, del reddito equivalente disponibile nel 2003 per provincia (Toscana).

Figura 4.3: Stima dei quartili, e del relativo errore standard, del reddito equivalente disponibile nel 2003 per provincia (Toscana).

Appendice A

Dettaglio delle tabelle del capitolo 4

In questa appendice si riporta il dettaglio dei dati presentati sotto forma di distribuzione nel capitolo 4.

Nelle seguenti tabelle si riporta l'errore relativo (RB), l'errore assoluto relativo (ARB), il tasso di copertura (CR), il vero errore quadratico medio e la sua stima bootstrap della stima dei quartili della variabile Y per piccola area della popolazione Ω_1 e della popolazione Ω_2, paragrafo 4.4.

A seguire si riporta l'errore relativo (RB), errore assoluto relativo (ARB), il tasso di copertura (CR), il vero errore quadratico medio e la sua stima bootstrap della stima dei quartili della variabile Y (reddito equivalente disponibile) per piccola area della "popolazione Albania", paragrafo 4.5[1].

In ogni tabella sarà specificato il quartile stimato, la popolazione a cui si riferisce la stima e il tipo di approccio bootstrap utilizzato: smooth non condizionato o empirico non condizionato.

Le ultime tabelle di questa appendice riguardano: **i.** la dimensione campionaria ($n_i, i = (1, \ldots, 36)$) e la dimensione della popolazione nelle piccole aree della "popolazione Albania" ($N_i, i = (1, \ldots, 36)$), usata nelle simulazioni design-based e **ii.** la dimensione campionaria riferita al numero delle famiglie selezionate nell'indagine EU-SILC del 2004 ($n_i, i = (1, \ldots, 10)$) e il numero totale di famiglie ($N_i, i = (1, \ldots, 10)$), ottenuto dal censimento Famiglie e Abitazioni del 2001, nelle province della regione Toscana.

[1] Con "popolazione Albania" si intende il campione LSMS Albania del 2003, come specificato nella simulazioni del paragrafo 4.5.

Tabella A.1: Stima del primo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.25)]$	$\widehat{RMSE}[\hat{q}_i(0.25)]$	$RB\%$	$ARB\%$	CR
1	0.72	0.71	−1.31	1.31	0.98
2	0.64	0.63	−2.10	2.10	0.94
3	0.68	0.65	−4.52	4.52	0.94
4	0.67	0.65	−2.17	2.17	0.94
5	0.66	0.60	−8.77	8.77	0.94
6	0.89	0.86	−3.37	3.37	0.93
7	0.56	0.59	4.15	4.15	0.96
8	0.57	0.58	2.84	2.84	0.95
9	0.72	0.68	−4.28	4.28	0.96
10	0.60	0.60	−0.14	0.14	0.96
11	0.42	0.43	0.36	0.36	0.96
12	0.56	0.54	−4.37	4.37	0.94
13	0.49	0.48	−2.81	2.81	0.95
14	0.49	0.47	−4.17	4.17	0.96
15	0.52	0.52	−0.75	0.75	0.95
16	0.47	0.43	−9.46	9.46	0.95
17	0.55	0.53	−4.37	4.37	0.94
18	0.41	0.41	0.92	0.92	0.97
19	0.48	0.45	−6.02	6.02	0.94
20	0.54	0.53	−2.15	2.15	0.95
21	0.39	0.39	−1.41	1.41	0.95
22	0.36	0.36	−0.09	0.09	0.96
23	0.36	0.37	2.78	2.78	0.96
24	0.37	0.35	−3.97	3.97	0.94
25	0.36	0.39	8.32	8.32	0.96
26	0.35	0.34	−2.59	2.59	0.95
27	0.39	0.38	−1.39	1.39	0.96
28	0.39	0.34	−11.48	11.48	0.92
29	0.38	0.38	−1.21	1.21	0.96
30	0.42	0.37	−12.28	12.28	0.90

Tabella A.2: Stima del secondo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.50)]$	$\widehat{RMSE}[\hat{q}_i(0.50)]$	$RB\%$	$ARB\%$	CR
1	0.66	0.65	−2.64	2.64	0.92
2	0.60	0.57	−5.59	5.59	0.95
3	0.61	0.60	−2.33	2.33	0.96
4	0.62	0.59	−4.60	4.60	0.95
5	0.59	0.55	−6.59	6.59	0.94
6	0.73	0.77	5.12	5.12	0.96
7	0.56	0.53	−5.24	5.24	0.95
8	0.53	0.54	0.79	0.79	0.95
9	0.61	0.62	2.18	2.18	0.98
10	0.58	0.55	−4.93	4.93	0.94
11	0.40	0.38	−3.52	3.52	0.92
12	0.50	0.49	−2.20	2.20	0.96
13	0.49	0.44	−9.77	9.77	0.93
14	0.44	0.43	−1.83	1.83	0.94
15	0.50	0.47	−6.31	6.31	0.94
16	0.40	0.39	−0.57	0.57	0.96
17	0.46	0.49	6.14	6.14	0.97
18	0.41	0.37	−9.00	9.00	0.93
19	0.42	0.42	−1.32	1.32	0.98
20	0.48	0.49	0.94	0.94	0.93
21	0.34	0.36	4.74	4.74	0.97
22	0.34	0.34	0.20	0.20	0.96
23	0.35	0.33	−6.32	6.32	0.93
24	0.35	0.32	−6.26	6.26	0.96
25	0.34	0.35	2.03	2.03	0.92
26	0.34	0.32	−7.36	7.36	0.92
27	0.35	0.35	−2.30	2.30	0.93
28	0.34	0.32	−7.18	7.18	0.92
29	0.36	0.34	−4.61	4.61	0.95
30	0.38	0.34	−11.37	11.37	0.92

Tabella A.3: Stima del terzo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.75)]$	$\widehat{RMSE}[\hat{q}_i(0.75)]$	$RB\%$	$ARB\%$	CR
1	0.72	0.70	−2.58	2.58	0.95
2	0.66	0.62	−5.93	5.93	0.94
3	0.70	0.65	−7.01	7.01	0.93
4	0.66	0.64	−3.27	3.27	0.94
5	0.68	0.60	−12.50	12.50	0.92
6	0.83	0.86	2.46	2.46	0.96
7	0.57	0.58	1.51	1.51	0.96
8	0.55	0.58	7.17	7.17	0.96
9	0.74	0.68	−7.74	7.74	0.94
10	0.61	0.60	−1.69	1.69	0.95
11	0.42	0.41	−0.18	0.18	0.94
12	0.51	0.54	5.19	5.19	0.96
13	0.54	0.48	−10.34	10.34	0.93
14	0.47	0.47	1.13	1.13	0.97
15	0.55	0.52	−5.11	5.11	0.94
16	0.42	0.43	0.81	0.81	0.94
17	0.50	0.53	6.15	6.15	0.96
18	0.43	0.40	−5.78	5.78	0.94
19	0.47	0.45	−4.18	4.18	0.96
20	0.54	0.53	−1.72	1.72	0.94
21	0.37	0.38	2.19	2.19	0.97
22	0.35	0.36	4.26	4.26	0.96
23	0.40	0.36	−10.09	10.09	0.92
24	0.35	0.35	−1.18	1.18	0.96
25	0.38	0.38	0.92	0.92	0.94
26	0.36	0.34	−6.18	6.18	0.94
27	0.36	0.37	2.41	2.41	0.96
28	0.36	0.34	−6.01	6.01	0.95
29	0.40	0.37	−8.18	8.18	0.92
30	0.41	0.36	−11.80	11.80	0.94

Tabella A.4: Stima del primo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.25)]$	$\widehat{RMSE}[\hat{q}_i(0.25)]$	$RB\%$	$ARB\%$	CR
1	0.72	0.71	−1.31	1.31	0.98
2	0.64	0.63	−2.10	2.10	0.94
3	0.68	0.65	−4.52	4.52	0.94
4	0.67	0.65	−2.17	2.17	0.94
5	0.66	0.60	−8.77	8.77	0.94
6	0.89	0.86	−3.37	3.37	0.93
7	0.56	0.59	4.15	4.15	0.96
8	0.57	0.58	2.84	2.84	0.95
9	0.72	0.68	−4.28	4.28	0.96
10	0.60	0.60	−0.14	0.14	0.96
11	0.42	0.43	0.36	0.36	0.96
12	0.56	0.54	−4.37	4.37	0.94
13	0.49	0.48	−2.81	2.81	0.95
14	0.49	0.47	−4.17	4.17	0.96
15	0.52	0.52	−0.75	0.75	0.95
16	0.47	0.43	−9.46	9.46	0.95
17	0.55	0.53	−4.37	4.37	0.94
18	0.41	0.41	0.92	0.92	0.97
19	0.48	0.45	−6.02	6.02	0.94
20	0.54	0.53	−2.15	2.15	0.95
21	0.39	0.39	−1.41	1.41	0.95
22	0.36	0.36	−0.09	0.09	0.96
23	0.36	0.37	2.78	2.78	0.96
24	0.37	0.35	−3.97	3.97	0.94
25	0.36	0.39	8.32	8.32	0.96
26	0.35	0.34	−2.59	2.59	0.95
27	0.39	0.38	−1.39	1.39	0.96
28	0.39	0.34	−11.48	11.48	0.92
29	0.38	0.38	−1.21	1.21	0.96
30	0.42	0.37	−12.28	12.28	0.90

Tabella A.5: Stima del secondo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.50)]$	$\widehat{RMSE}[\hat{q}_i(0.50)]$	$RB\%$	$ARB\%$	CR
1	0.68	0.65	−4.74	4.74	0.96
2	0.57	0.56	−1.12	1.12	0.96
3	0.63	0.59	−6.56	6.56	0.94
4	0.64	0.59	−8.04	8.04	0.94
5	0.55	0.55	0.32	0.32	0.97
6	0.76	0.78	1.76	1.76	0.95
7	0.52	0.53	1.08	1.08	0.95
8	0.53	0.53	−0.72	0.72	0.94
9	0.70	0.62	−11.90	11.90	0.92
10	0.59	0.55	−7.54	7.54	0.94
11	0.41	0.39	−5.67	5.67	0.93
12	0.50	0.49	−2.05	2.05	0.94
13	0.46	0.43	−5.18	5.18	0.96
14	0.43	0.43	−0.04	0.04	0.94
15	0.46	0.47	2.30	2.30	0.96
16	0.42	0.39	−7.03	7.03	0.93
17	0.52	0.48	−7.11	7.11	0.95
18	0.37	0.37	1.90	1.90	0.95
19	0.43	0.41	−3.12	3.12	0.94
20	0.48	0.48	1.51	1.51	0.97
21	0.35	0.35	0.62	0.62	0.95
22	0.33	0.33	0.65	0.65	0.94
23	0.34	0.33	−1.30	1.30	0.94
24	0.31	0.32	4.17	4.17	0.96
25	0.34	0.35	2.57	2.57	0.95
26	0.32	0.32	−1.70	1.70	0.92
27	0.34	0.35	2.15	2.15	0.98
28	0.34	0.32	−7.91	7.91	0.93
29	0.37	0.34	−5.80	5.80	0.94
30	0.37	0.34	−9.03	9.03	0.93

Tabella A.6: Stima del terzo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_1 (errori Normali), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.75)]$	$\widehat{RMSE}[\hat{q}_i(0.75)]$	$RB\%$	$ARB\%$	CR
1	0.74	0.70	−4.54	4.54	0.93
2	0.61	0.62	0.63	0.63	0.97
3	0.69	0.65	−5.55	5.55	0.94
4	0.68	0.64	−5.87	5.87	0.96
5	0.63	0.60	−3.52	3.52	0.94
6	0.91	0.87	−4.68	4.68	0.93
7	0.60	0.58	−3.58	3.58	0.97
8	0.58	0.58	−1.21	1.21	0.94
9	0.72	0.68	−5.41	5.41	0.94
10	0.65	0.60	−7.68	7.68	0.93
11	0.46	0.42	−8.16	8.16	0.95
12	0.56	0.53	−4.64	4.64	0.93
13	0.50	0.47	−5.44	5.44	0.93
14	0.48	0.47	−2.85	2.85	0.97
15	0.51	0.52	1.26	1.26	0.97
16	0.46	0.42	−7.50	7.50	0.95
17	0.61	0.53	−13.34	13.34	0.92
18	0.40	0.40	0.04	0.04	0.95
19	0.49	0.45	−8.73	8.73	0.95
20	0.52	0.52	0.72	0.72	0.94
21	0.38	0.38	1.48	1.48	0.93
22	0.35	0.36	2.70	2.70	0.95
23	0.36	0.36	−1.05	1.05	0.95
24	0.35	0.35	−0.62	0.62	0.95
25	0.40	0.38	−6.40	6.40	0.93
26	0.36	0.34	−4.84	4.84	0.95
27	0.38	0.37	−2.18	2.18	0.98
28	0.34	0.34	−0.34	0.34	0.96
29	0.40	0.37	−7.02	7.02	0.95
30	0.38	0.36	−5.09	5.09	0.94

Tabella A.7: Stima del primo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.25)]$	$\widehat{RMSE}[\hat{q}_i(0.25)]$	$RB\%$	$ARB\%$	CR
1	0.65	0.67	2.99	2.99	0.98
2	0.59	0.57	−3.41	3.41	0.96
3	0.65	0.61	−6.36	6.36	0.95
4	0.68	0.60	−12.43	12.43	0.92
5	0.53	0.57	6.40	6.40	0.96
6	0.92	0.79	−14.19	14.19	0.91
7	0.51	0.54	5.38	5.38	0.97
8	0.52	0.53	2.77	2.77	0.96
9	0.63	0.64	1.50	1.50	0.96
10	0.54	0.56	2.70	2.70	0.98
11	0.38	0.38	0.18	0.18	0.93
12	0.46	0.49	7.93	7.93	0.96
13	0.45	0.44	−1.70	1.70	0.96
14	0.40	0.43	6.95	6.95	0.96
15	0.48	0.48	−1.16	1.16	0.94
16	0.39	0.40	1.74	1.74	0.96
17	0.52	0.49	−5.94	5.94	0.96
18	0.37	0.37	1.28	1.28	0.95
19	0.42	0.41	−1.25	1.25	0.95
20	0.52	0.49	−6.37	6.37	0.94
21	0.35	0.35	1.03	1.03	0.95
22	0.34	0.33	−3.60	3.60	0.95
23	0.33	0.33	0.57	0.57	0.96
24	0.32	0.32	2.02	2.02	0.97
25	0.37	0.35	−6.66	6.66	0.94
26	0.30	0.32	4.60	4.60	0.97
27	0.38	0.34	−8.79	8.79	0.95
28	0.31	0.32	1.19	1.19	0.94
29	0.32	0.34	7.19	7.19	0.98
30	0.35	0.34	−2.99	2.99	0.94

Tabella A.8: Stima del secondo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.50)]$	$\widehat{RMSE}[\hat{q}_i(0.50)]$	$RB\%$	$ARB\%$	CR
1	0.81	0.83	2.44	2.44	0.96
2	0.74	0.72	−2.46	2.46	0.96
3	0.88	0.76	−13.57	13.57	0.92
4	0.90	0.74	−17.60	17.60	0.92
5	0.69	0.71	2.03	2.03	0.95
6	1.13	1.00	−11.23	11.23	0.94
7	0.70	0.67	−3.53	3.53	0.94
8	0.69	0.67	−2.77	2.77	0.97
9	0.80	0.80	−0.57	0.57	0.96
10	0.67	0.70	5.51	5.51	0.96
11	0.55	0.48	−12.43	12.43	0.94
12	0.60	0.62	4.06	4.06	0.96
13	0.57	0.55	−4.05	4.05	0.94
14	0.50	0.55	9.28	9.28	0.97
15	0.65	0.61	−6.27	6.27	0.92
16	0.57	0.49	−12.98	12.98	0.94
17	0.63	0.62	−2.89	2.89	0.94
18	0.49	0.47	−4.92	4.92	0.96
19	0.50	0.53	4.34	4.34	0.96
20	0.63	0.61	−3.50	3.50	0.96
21	0.40	0.44	10.87	10.87	0.98
22	0.45	0.42	−6.79	6.79	0.92
23	0.42	0.42	−1.21	1.21	0.94
24	0.42	0.40	−4.01	4.01	0.93
25	0.47	0.44	−7.11	7.11	0.94
26	0.39	0.40	1.26	1.26	0.97
27	0.48	0.44	−9.14	9.14	0.93
28	0.42	0.40	−4.68	4.68	0.94
29	0.44	0.43	−2.25	2.25	0.94
30	0.44	0.42	−4.79	4.79	0.94

Tabella A.9: Stima del terzo quartile. Approccio bootstrap empirico non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.75)]$	$\widehat{RMSE}[\hat{q}_i(0.75)]$	$RB\%$	$ARB\%$	CR
1	1.25	1.25	0.30	0.30	0.94
2	1.17	1.08	−7.55	7.55	0.94
3	1.16	1.12	−3.14	3.14	0.94
4	1.29	1.10	−14.97	14.97	0.94
5	1.06	1.06	−0.78	0.78	0.96
6	1.69	1.60	−5.42	5.42	0.94
7	1.04	1.02	−1.78	1.78	0.94
8	1.09	1.01	−7.62	7.62	0.94
9	1.17	1.23	4.89	4.89	0.96
10	1.05	1.05	0.50	0.50	0.96
11	0.82	0.72	−12.15	12.15	0.90
12	0.93	0.93	0.72	0.72	0.97
13	0.85	0.83	−2.24	2.24	0.96
14	0.81	0.82	1.36	1.36	0.94
15	0.93	0.91	−2.83	2.83	0.95
16	0.80	0.73	−8.44	8.44	0.93
17	0.96	0.93	−3.38	3.38	0.94
18	0.68	0.70	4.11	4.11	0.98
19	0.77	0.79	2.40	2.40	0.94
20	0.94	0.92	−1.60	1.60	0.95
21	0.63	0.66	4.65	4.65	0.97
22	0.58	0.63	8.41	8.41	0.96
23	0.61	0.62	2.87	2.87	0.96
24	0.63	0.60	−4.38	4.38	0.94
25	0.68	0.66	−2.60	2.60	0.94
26	0.58	0.59	1.65	1.65	0.95
27	0.73	0.65	−10.48	10.48	0.93
28	0.56	0.59	6.28	6.28	0.98
29	0.68	0.64	−6.41	6.41	0.92
30	0.65	0.63	−2.82	2.82	0.96

Tabella A.10: Stima del primo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.25)]$	$\widehat{RMSE}[\hat{q}_i(0.25)]$	$RB\%$	$ARB\%$	CR
1	0.64	0.66	2.78	2.78	0.96
2	0.56	0.57	2.14	2.14	0.96
3	0.61	0.61	−0.30	0.30	0.96
4	0.58	0.60	4.10	4.10	0.96
5	0.57	0.56	−2.07	2.07	0.94
6	0.92	0.80	−13.51	13.51	0.92
7	0.54	0.53	−0.94	0.94	0.96
8	0.58	0.53	−8.97	8.97	0.94
9	0.65	0.63	−3.48	3.48	0.96
10	0.64	0.56	−11.98	11.98	0.94
11	0.39	0.38	−2.99	2.99	0.96
12	0.45	0.49	8.84	8.84	0.97
13	0.47	0.44	−5.86	5.86	0.92
14	0.47	0.43	−7.62	7.62	0.92
15	0.46	0.47	2.98	2.98	0.96
16	0.39	0.40	0.47	0.47	0.96
17	0.50	0.49	−2.43	2.43	0.94
18	0.41	0.37	−9.26	9.26	0.92
19	0.42	0.41	−2.62	2.62	0.96
20	0.51	0.50	−2.21	2.21	0.95
21	0.38	0.35	−8.24	8.24	0.92
22	0.33	0.33	0.66	0.66	0.96
23	0.35	0.33	−3.94	3.94	0.95
24	0.30	0.32	8.60	8.60	0.97
25	0.35	0.35	−0.31	0.31	0.94
26	0.32	0.31	−2.51	2.51	0.96
27	0.36	0.35	−4.17	4.17	0.95
28	0.29	0.31	8.78	8.78	0.98
29	0.36	0.34	−4.23	4.23	0.94
30	0.36	0.33	−6.06	6.06	0.94

Tabella A.11: Stima del secondo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.50)]$	$\widehat{RMSE}[\hat{q}_i(0.50)]$	$RB\%$	$ARB\%$	CR
1	0.76	0.82	8.31	8.31	0.95
2	0.75	0.72	−4.40	4.40	0.94
3	0.76	0.76	−0.05	0.05	0.95
4	0.74	0.75	0.90	0.90	0.96
5	0.70	0.70	0.96	0.96	0.94
6	1.09	1.00	−7.89	7.89	0.94
7	0.72	0.67	−5.67	5.67	0.95
8	0.80	0.68	−14.95	14.95	0.91
9	0.81	0.80	−1.68	1.68	0.96
10	0.78	0.71	−8.71	8.71	0.94
11	0.48	0.48	0.21	0.21	0.94
12	0.59	0.62	5.82	5.82	0.98
13	0.59	0.56	−4.80	4.80	0.94
14	0.58	0.55	−5.32	5.32	0.94
15	0.57	0.60	5.31	5.31	0.96
16	0.54	0.50	−7.21	7.21	0.94
17	0.61	0.62	1.20	1.20	0.97
18	0.55	0.47	−15.37	15.37	0.91
19	0.57	0.52	−7.94	7.94	0.92
20	0.58	0.62	6.54	6.54	0.96
21	0.46	0.44	−2.95	2.95	0.95
22	0.46	0.42	−10.01	10.01	0.94
23	0.45	0.42	−7.14	7.14	0.94
24	0.40	0.40	1.84	1.84	0.95
25	0.44	0.44	−0.30	0.30	0.97
26	0.38	0.40	5.18	5.18	0.98
27	0.42	0.44	4.68	4.68	0.98
28	0.38	0.40	3.11	3.11	0.94
29	0.44	0.43	−2.42	2.42	0.96
30	0.44	0.42	−3.82	3.82	0.94

Tabella A.12: Stima del terzo quartile. Approccio bootstrap smooth non condizionato. Popolazione Ω_2 (errori χ^2), paragrafo 4.4.

Area	$RMSE[\hat{q}_i(0.75)]$	$\widehat{RMSE}[\hat{q}_i(0.75)]$	$RB\%$	$ARB\%$	CR
1	1.25	1.26	0.53	0.53	0.96
2	1.05	1.09	3.20	3.20	0.97
3	1.16	1.11	−4.03	4.03	0.94
4	1.09	1.12	3.23	3.23	0.96
5	1.00	1.06	6.27	6.27	0.96
6	1.73	1.58	−8.56	8.56	0.92
7	1.04	1.03	−0.27	0.27	0.96
8	1.11	1.04	−6.66	6.66	0.96
9	1.18	1.22	3.21	3.21	0.98
10	1.03	1.07	3.49	3.49	0.96
11	0.70	0.73	3.75	3.75	0.96
12	0.82	0.94	14.79	14.79	0.98
13	0.84	0.85	0.82	0.82	0.96
14	0.92	0.84	−9.10	9.10	0.94
15	0.95	0.90	−5.27	5.27	0.95
16	0.72	0.74	3.15	3.15	0.96
17	1.02	0.94	−7.87	7.87	0.94
18	0.85	0.71	−16.44	16.44	0.90
19	0.84	0.79	−6.26	6.26	0.94
20	0.99	0.94	−5.22	5.22	0.96
21	0.68	0.67	−1.40	1.40	0.94
22	0.63	0.63	−0.04	0.04	0.97
23	0.65	0.63	−2.32	2.32	0.96
24	0.59	0.60	1.87	1.87	0.96
25	0.71	0.67	−6.25	6.25	0.96
26	0.63	0.60	−5.39	5.39	0.93
27	0.63	0.67	5.16	5.16	0.96
28	0.62	0.59	−4.29	4.29	0.95
29	0.69	0.65	−5.50	5.50	0.94
30	0.67	0.64	−5.13	5.13	0.92

Tabella A.13: Stima del primo quartile. Approccio bootstrap smooth non condizionato. "Popolazione Albania", paragrafo 4.5.

Area	$RMSE[\hat{q}_i(0.25)]$	$\widehat{RMSE}[\hat{q}_i(0.25)]$	$RB\%$	$ARB\%$	CR
1	1053.45	967.19	−8.19	8.19	0.96
2	431.59	900.94	108.75	108.75	1.00
3	2400.84	1446.44	−39.75	39.75	0.78
4	2636.39	1395.85	−47.05	47.05	0.58
5	734.01	646.38	−11.94	11.94	0.92
6	790.27	813.80	2.98	2.98	0.95
7	671.57	819.39	22.01	22.01	1.00
8	737.69	667.94	−9.46	9.46	0.95
9	897.47	967.04	7.75	7.75	0.96
10	2234.51	1552.97	−30.50	30.50	0.88
11	1842.30	1577.78	−14.36	14.36	0.92
12	1176.77	1117.58	−5.03	5.03	0.95
13	594.47	1125.87	89.39	89.39	1.00
14	855.08	838.10	−1.99	1.99	0.93
15	1613.70	1596.31	−1.08	1.08	0.96
16	1751.01	1537.45	−12.20	12.20	0.94
17	515.68	735.60	42.65	42.65	0.98
18	716.93	1355.26	89.04	89.04	1.00
19	1875.45	1362.13	−27.37	27.37	0.90
20	676.16	713.76	5.56	5.56	0.96
21	967.99	808.50	−16.48	16.48	0.92
22	1379.46	1515.39	9.85	9.85	0.94
23	2152.00	1531.13	−28.85	28.85	0.86
24	1160.31	1562.59	34.67	34.67	1.00
25	1351.52	1434.52	6.14	6.14	0.90
26	1189.54	1485.68	24.90	24.90	0.98
27	2119.89	1460.64	−31.10	31.10	0.85
28	2142.81	1609.19	−24.90	24.90	0.86
29	866.78	1467.84	69.34	69.34	1.00
30	1908.73	1588.95	−16.75	16.75	0.92
31	2198.79	1430.25	−34.95	34.95	0.75
32	830.78	850.34	2.36	2.36	0.97
33	1334.43	1545.88	15.85	15.85	0.96
34	360.97	374.31	3.70	3.70	0.95
35	1057.79	1110.66	5.00	5.00	0.99
36	978.27	798.39	−18.39	18.39	0.88

Tabella A.14: Stima del secondo quartile. Approccio bootstrap smooth non condizionato. "Popolazione Albania", paragrafo 4.5.

Area	$RMSE[\hat{q}_i(0.50)]$	$\widehat{RMSE}[\hat{q}_i(0.50)]$	$RB\%$	$ARB\%$	CR
1	1094.57	1109.20	1.34	1.34	0.97
2	563.07	1197.58	112.69	112.69	1.00
3	2722.56	1725.52	−36.62	36.62	0.70
4	1998.65	1732.67	−13.31	13.31	0.92
5	881.24	883.64	0.27	0.27	0.96
6	1294.15	1005.17	−22.33	22.33	0.86
7	955.03	1132.00	18.53	18.53	0.98
8	873.41	857.75	−1.79	1.79	0.96
9	973.93	1143.15	17.38	17.38	0.97
10	2184.33	1922.02	−12.01	12.01	0.92
11	1752.58	1916.02	9.33	9.33	0.96
12	1465.58	1413.93	−3.52	3.52	0.96
13	363.31	1419.98	290.85	290.85	1.00
14	1183.97	1049.10	−11.39	11.39	0.94
15	1654.47	1969.45	19.04	19.04	0.96
16	2146.15	1863.88	−13.15	13.15	0.85
17	988.64	968.21	−2.07	2.07	0.95
18	1034.12	1594.51	54.19	54.19	1.00
19	2792.98	1505.25	−46.11	46.11	0.77
20	818.25	895.79	9.48	9.48	0.98
21	1341.45	1082.77	−19.28	19.28	0.90
22	1716.52	1859.24	8.31	8.31	0.98
23	2782.96	1873.41	−32.68	32.68	0.87
24	1450.44	1892.43	30.47	30.47	0.93
25	1932.43	1761.10	−8.87	8.87	0.99
26	1968.39	1820.38	−7.52	7.52	0.92
27	2422.83	1734.48	−28.41	28.41	0.87
28	2203.23	1938.56	−12.01	12.01	0.95
29	831.62	1804.43	116.98	116.98	1.00
30	2053.67	1947.23	−5.18	5.18	0.92
31	2082.56	1687.78	−18.96	18.96	0.86
32	1029.68	1056.52	2.61	2.61	0.96
33	2456.82	1902.18	−22.58	22.58	0.90
34	645.15	502.12	−22.17	22.17	0.88
35	892.19	1448.54	62.36	62.36	1.00
36	1451.62	1087.18	−25.11	25.11	0.86

Tabella A.15: Stima del terzo quartile. Approccio bootstrap smooth non condizionato. "Popolazione Albania", paragrafo 4.5.

Area	$RMSE[\hat{q}_i(0.75)]$	$\widehat{RMSE}[\hat{q}_i(0.75)]$	$RB\%$	$ARB\%$	CR
1	2014.46	1764.04	−12.43	12.43	0.93
2	1281.34	1750.19	36.59	36.59	0.99
3	3238.03	2368.12	−26.87	26.87	0.84
4	1798.68	2396.02	33.21	33.21	1.00
5	1139.27	1370.63	20.31	20.31	0.98
6	1475.10	1559.94	5.75	5.75	0.97
7	1766.06	1837.98	4.07	4.07	0.92
8	1826.75	1427.06	−21.88	21.88	0.84
9	1423.73	1746.72	22.69	22.69	0.98
10	2364.04	2716.71	14.92	14.92	0.98
11	2143.03	2679.20	25.02	25.02	0.96
12	2159.46	2070.29	−4.13	4.13	0.93
13	1981.82	1869.25	−5.68	5.68	1.00
14	2052.35	1744.25	−15.01	15.01	0.91
15	2297.03	2767.12	20.47	20.47	0.97
16	2313.67	2668.01	15.32	15.32	1.00
17	1476.68	1522.21	3.08	3.08	0.97
18	1603.89	2738.88	70.76	70.76	1.00
19	3806.41	2772.46	−27.16	27.16	0.84
20	1360.32	1378.52	1.34	1.34	0.97
21	1623.19	1753.21	8.01	8.01	0.97
22	2255.79	2654.85	17.69	17.69	1.00
23	3908.65	2699.51	−30.94	30.94	0.84
24	1932.12	2636.31	36.45	36.45	0.98
25	2018.66	2550.07	26.33	26.33	1.00
26	3084.10	2652.76	−13.99	13.99	0.96
27	2923.62	2392.03	−18.18	18.18	0.92
28	2376.97	2683.77	12.91	12.91	0.96
29	854.67	2474.87	189.57	189.57	1.00
30	2711.80	2713.87	0.08	0.08	0.96
31	1747.32	2479.25	41.89	41.89	0.96
32	1760.63	1742.26	−1.04	1.04	0.97
33	3054.57	2688.09	−12.00	12.00	0.94
34	884.02	765.08	−13.45	13.45	0.93
35	1466.89	2090.97	42.54	42.54	1.00
36	2209.91	1767.46	−20.02	20.02	0.91

Tabella A.16: Dimensione nelle piccola aree della "popolazione Albania" e dimensione campionaria utilizzata nella simulazione design-based. Unità di riferimento: nucleo familiare (household).

	N_i	n_i
1	120	12
2	128	13
3	16	5
4	16	5
5	232	23
6	160	16
7	152	15
8	224	22
9	120	12
10	32	5
11	48	5
12	88	9
13	8	5
14	136	14
15	38	5
16	32	5
17	184	18
18	64	6
19	64	6
20	200	20
21	152	15
22	24	5
23	32	5
24	32	5
25	16	5
26	24	5
27	16	5
28	48	5
29	23	5
30	48	5
31	16	5
32	138	14
33	32	5
34	688	69
35	88	9
36	152	15

Tabella A.17: Numero delle famiglie nelle province della regione Toscana (fonte: censimento Famiglie e Abitazioni del 2001) e numero delle famiglie campionate nell'indagine EU-SILC del 2004 nelle province della regione Toscana.

Provincia	N_i	n_i
Massa-Carrara	80811	126
Lucca	146118	123
Pistoia	104467	137
Firenze	376255	545
Livorno	133730	142
Pisa	150259	158
Arezzo	123881	161
Siena	101400	135
Grosseto	87721	70
Prato	83618	154

Bibliografia

Aigner D.; Amemiya T.; Poirier D. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, **17**, 377–96.

Andersen R. (2008). *Modern Methods for Robust Regression*. Sage Pubblications.

Aragon Y.; Casanova S.; Chambers R.; Leoconte E. (2006). Conditional ordering using nonparametric expectiles. In corso di pubblicazione.

Battese, G.E. e. a. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Bickel P.; Freedman D. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9**(6), 1196–217.

Bogue D. (1950). A technique for making extensive postcensal estimates. *Journal of the American Statistical Association*, **45**, 149–63.

Bogue D.; Duncan B. (1959). A composite method of estimating postcensal population of small areas by age, sex and colour. *Vital Statistics*, **Special Report 47**(6).

Bowman A.; Hall P.; Prvan T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, **85**, 799–808.

Brackstone G. (1987). Small area data: Policy issues and technical challenges In *Small Area Statistics*. A cura di Platek R., Rao J., Sarndal C., Singh M., pp. 3–20. Wiley, New York.

Breckling J.; Chambers R. (1988). M-quantiles. *Biometrika*, **75**(4), 761–771.

Carpenter J.; Bithell J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, **19**, 1141–64.

Ceccarelli C.; Di Marco M.; Rinaldelli C. (2008). L'indagine europea sui redditi e le condizioni di vita delle famiglie (eu-silc). *Metodi e Norme*, **37**.

Chambers R. (2005). Calibrated weighting for small area estimation. Relazione tecnica, Southampton statistical Sciences Research Institute. Disponibile all'indirizzo http://eprints.soton.ac.uk/14074/.

Chambers R.; Dunstan (1986). Estimating distribution function from survey data. *Biometrika*, **73**, 597–604.

Chambers R.; Tzavidis N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**(2), 255–68.

Chambers R.; Dorfman A.; Peter H. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, **79**(3), 577–582.

Chao M. (1982). A general purpose unequal probability sampling plan. *Biometrika*, **69**, 652–6.

Chiandotto B. (1996). L'informazione statistica a livello territoriale: significatività, problemi e limiti. In *Terza Conferenza Nazionale di Statistica.*, pp. 24–26, Roma.

Datta G.; Gosh M. (1991). Bayesian prediction in linear model: Application to small area estimation. *Annals of Statistics*, **19**, 1784–70.

Datta G.; Lahiri P. (1997). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613–627.

Davison A.; Hinkley D. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

Demidenko E. (2004). *Mixed models, Theory and Applications*. Jhon Wiley, Nwe York.

Efron B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Efron B.; Tibshirani R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Ericksen E. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, **69**(348), 867–75.

Falorsi P.; Falorsi S.; Russo A. (1994). Empirical comparision of small area estimation methods for the italian labour force survey. *Survey Methodology*, **20**(2), 171–176.

Fay R.; Herriot R. (1979). Estimation of icome from small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–77.

Giommi A.; Rocco E. (2003). Model assisted small area estimation in a municipal labour force survey. In *International Statistical Institute*, 54th Session, Berlino.

Goldstein H. (2003). *Multilevel Statistical Model*. Arnold, London.

Gonzalez M. (1973). Use and evaluaion of synthetic estimates. In *Proceedings of the Social Statistics Section*, pp. 33–36. American Statistical Association.

Gori E.; Marchetti G. (1987). Problemi e metodi di nanalisi di dati statistici per piccole aree. In *Atti del Convegno su Informazione ed Analisi Statistica per Aree Regionali e Sub-Regionali.*, pp. 47–71. SIS.

Gosh M.; Rao J. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, **9**(1), 55–93.

Green W. (2000). *Econometric Analysis*. Prentice-Hall.

Grosh M.; Glewwe P. (1995). A guide to living standard measurement study surveys and their data sets. *The World Bank, LSMS Working Paper*, **-**(120).

Hall P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

He X. (1997). Quantile curves without crossing. *American Statistician*, **51**, 186–192.

Hoaglin D.; Mosteller F.; Tukey J. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley.

Horvitz D.; Thompson D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–85.

Huber P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Huber P. (1981). *Robust Statistics*. Wiley.

Jiang J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. Springer Verlag, New York.

Koenker R. (2005). *Quantile regression*. Cambridge University Press.

Koenker R.; Bassett G. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Kokic P.; Chambers R.; Breckling J.; Beare S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, **15**(4), 445–51.

Kuk A. (1988). Estimation of distribution function and medians under sampling with unequal probabilities. *Biometrika*, **75**, 97–103.

Li Q.; Racine J. (2007). *Nonparametric econometric: theory and practice.* Princeton University Press.

Lombardia M.; Gonzalez-Manteiga W.; Prada-Sanchez J. (2003). Bootstrapping the chambers-dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367–388.

Madow W. (1949). On the thoeory of systematic sampling ii. *Annals of Mathematical Statistics*, **20**, 333–54.

McCullogh P.; Searle S. (2001). *Generalized, Linear and Mixed Models*. Jhon Wiley, Nwe York.

Mood A.; Graybill F.; Boes D. (1974). *Introduction to the theory of statistics*. McGraw-Hill.

Mosteller F.; Tukey J. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley.

Newey W.; Powell J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**(4), 819–47.

of the Census U. B. (1966). *The Census Component Method.*

Pfeffermann D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, **70**(1), 125–43.

Purcell N.; Kish L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, **48**, 3–18.

Rao J. (2003). *Small Area Estimation*. New York:Wiley.

Rao J.; Kovar J.; Mantel H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, **77**(2), 365–75.

Richardson A.; Welsh A. (1996). Covariate screening in mixed linear models. *Journal of Multivariate Analysis*, **58**, 27–54.

Rousseeuw P. J.; Leroy A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Royall R.; Cumberland W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, **73**, 351–58.

Russo A. (1996). Stimatori per piccole aree: problemi aperti. In *Atti del Convegno su 100 anni di indagini campionarie.*, pp. 287–311, Roma. CISU.

Salvati N. (2004). *La correlazione spaziale nella stima per piccole aree: metodi proposti e casi di studio.* Tesi di Dottorato di Ricerca, Dottorato in Statistica Applicata - XVI.

Särndal C. (1993). Panel discussion in: Small area statistics and survey design. In *International Scientific Conference*, Varsavia. Central Statistical Office.

Särndal C.; Swensson B.; Wretman J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Shaible W. (1978). Chosing weights for composite estimators for small area statistics. In *Proceedings of the Section on Survey Research Methods.*, pp. 741–46. American Statistical Association.

Shao J.; Tu D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Singh A.; Gambino J.; Mantel H. (1994). Issues and strategies for small area data. *Survey Methodology*, **20**(1), 3–22.

Tzavidis N.; Salvati N.; Pratesi M.; Chambers R. (2008a). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, **17**(3).

Tzavidis N.; Marchetti S.; Chambers R. (2008b). Robust estimation of small area means and quantiles. In corso di pubblicaziones.

Wang S.; Dorfman A. (1996). A new estimator for the finite population distribution function. *Biometrika*, **83**, 639–52.

Wasserman L. (2004). *All of Statistics: a concise course in statistical inference*. Springer, New York.

Wilcox R. (2003). *Applying contemporary statistical techniques*. Academic Press.

Woodruff R. (1952). Confidence intervals for medians and others position measures. *Journal of the American Statistical Association*, **66**, 411–14.

Wu C.; Sitter R. (2001). Variance estimator for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, **29**.