Università degli Studi di Firenze
Dipartimento di Statistica “G. Parenti”
Dottorato di Ricerca in Statistica Applicata
XXII ciclo - SECS-S/01

Contatti sociali, stima, ed inferenza su parametri di infezioni a trasmissione diretta. Il caso della varicella.

Luigi Marangi

Tutor: **Prof. Piero Manfredi**
Co-tutor: **Prof. Monica Pratesi**
Coordinatore:**Fabio Corradi**

SOMMARIO

Indice delle figure

Indice delle tabelle

1 Introduzione

1.1 Motivazioni: stima di parametri di trasmissione di infezioni

Molte malattie, quali il morbillo, la parotite, la rosolia, l'influenza e la varicella, sono trasmesse per via respiratoria o contatto diretto. I *contatti sociali* sono quindi fattori critici per spiegare le dinamiche di *trasmissione* di un gran numero di agenti infettivi. La conoscenza dei comportamenti di contatto è quindi fondamentale per elaborare strategie di contenimento di un nuovo agente infettivo potenzialmente devastante come un virus pandemico (Wallinga et al., 2006), o per progettare efficaci misure di controllo per le infezioni endemiche, come per esempio la poliomielite oppure il morbillo. In particolare, è importante identificare in una popolazione le coorti d'età a cui deve essere mirata la vaccinazione. Nonostante la rilevanza del tema, la conoscenza dei meccanismi di contatto alla base della diffusione di malattie infettive a contatto diretto è ancora limitata. La struttura dei contatti nei modelli matematici di infezioni a contatto diretto (Hethcote,1989) è stata fin a tempi molto recenti stimata per via *indiretta*. Per esempio, in popolazioni stratificate per età, i tassi di trasmissione tra gruppi di età, che costituiscono la matrice 'Who-Aquires-Infection-From-Whom' (WAIFW, Anderson e May, 1991) sono stati tradizionalmente stimati calibrando i modelli ai dati epidemiologici (tipicamente rappresentati da una stima della forza dell'infezione al variare dell'età), sotto ipotesi a priori che permettono di ridurre il numero di parametri ignoti al fine di renderli stimabili. Per esempio nelle prime fasi della modellistica epidemiologica è stata utilizzata esclusivamente l'ipotesi di mixing omogeneo, equivalente ad assumere una popolazione composta da individui identici nei loro comportamenti sociali. Successivamente sono state introdotte, con particolare riferimento alla trasmissione delle infezioni esantematiche infantili, varie strategie (per esempio le ipotesi di Mixing proporzionale, preferenziale di Hethcote (1996), oppure le varie forme di matrici WAIFW di Anderson e May, 1991) per tenere conto della evidente eterogeneità nei comportamenti sociali, dominati dal picco di contatto sociali nella fase scolare. L'idea di base è quella di imporre appropriate configurazioni di 'contatti sociali' definite a priori ed appropriatamente ristrette, al fine di ridurre la dimensione dello spazio parametrico. Altri studi legati rilevanti sono Greenhalgh and Dietz (1994), Farrington et al. (2001), Van Effelterre et al. (2009), Farrington e Whitaker (2005).

1.2 Approcci diretti alla stima dei contatti sociali

In tempi recenti si sono resi disponibili in alcuni paesi Europei dati sui contatti sociali a rischio per la trasmissione delle infezioni a trasmissione diretta (progetto "Polymod", Mossong et al., 2008).
In effetti, il primo lavoro apparso in letteratura sull'utilizzazione di dati di contatto per età al fine di stimare parametri di trasmissione è due anni antecedente (Wallinga et al., 2006). Sebbene questo ultimo lavoro fornisca la metodologia innovativa di stima dei parametri di trasmissione che viene utilizzata in questa tesi, la struttura dei dati che gli autori utilizzano è relativamente semplice (si riportano soltanto i numeri delle conversazioni che i partecipanti hanno avuto con persone di differenti età durante una settimana tipica) e non permettono di investigare sul ruolo delle diverse tipologie di contatto (intimo, a casa, a scuola,a lavoro) nella diffusione dell'infezione.
Lo studio Polymod rappresenta uno studio di tipo campionario, armonizzato, svolto in diversi paesi europei, che estende significativamente l'approccio di Wallinga et al., 2006. L'approccio utilizzato nel progetto Polymod è un approccio "*diretto*" allo studio dei contatti sociali. Lo scopo degli approcci diretti è quello di superare le numerose limitazioni degli approcci indiretti sopra descritti. L'idea di base è quella di "creare" la *matrice dei contatti per età* ricavandola "direttamente" dai dati sui contatti sociali. A tale scopo occorre definire cosa si intende per 'contatto a rischio' rilevante per la trasmissione dell'infezione, per esempio una conversazione ravvicinata tra due soggetti senza interposizione di barriere, oppure un contatto fisico tra una madre e i suoi figli, e raccogliere dati da

indagini campionarie a tal fine. Nell'indagine Polymod i dati sui contatti sono stati raccolti mediante "diari" di contatto: ogni partecipante riceve un diario in cui deve registrare tutti i contatti avuti – in base ad appropriate definizioni contatto a rischio - in un giorno della settimana assegnato casualmente.
I dati Polymod offrono delle prospettive del tutto nuove nella stima dei parametri di trasmissione delle infezioni. Infatti tramite dati "diretti" di contatto si forniscono delle stime per le matrici di contatto per età, offrendo una soluzione al problema di sottoidentificazione tipico dei modelli epidemiologici. Lo studio Polymod ha anche stimolato la produzione di tecniche alternative di stima di dati di contatto, come i dati di utilizzo del tempo, o "Time Use" (Zagheni et al., 2008), oppure dati di tipo simulativo (Del Valle et al., 2007, Iozzi et al., PLoS Computational Biology, 2010).

1.3 Il fuoco della tesi: stima e inferenza su coefficienti di trasmissione

Questa tesi si occupa di una serie di problemi legati alla stima, ed all'inferenza, sui parametri di trasmissione di infezioni, per una varietà di tipologie di matrici di contatti sociali per età (contatti totali, per prossimità, per durata, per luogo), come disponibili dall'indagine Polymod per l'Italia. In particolare ci serviamo della varicella come caso di studio, per motivi di carattere tecnico ossia il fatto che essa sia a tutt'oggi una infezione nella sua fase pre-vaccinale. Questo consente di semplificare la struttura del modello matematico di adattamento ai dati, ipotizzando una situazione di equilibrio pre-vaccinale non perturbata dalla vaccinazione.
In presenza di dati di contatto – come i dati Polymod - il problema della stima della trasmissione è ricondotto alla stima di un piccolo numero di *coefficienti di trasmissione* (indicati con il simbolo q in questa tesi) specifici per tipologia di contatto. Tali coefficienti possono essere stimati mediante adattamento di un opportuno modello di regressione ai dati di sierologia, per mezzo di tecniche di verosimiglianza oppure di adattamento nonlineare. La costruzione del modello statistico si basa sul modello matematico Suscettibile-Infettivo-Rimosso per la trasmissione delle infezioni nel suo regime di equilibrio. I dati di sierologia sono la più comune fonte di informazione sull'esperienza di infezione: sono dati campionari tipicamente di tipo trasversale, stratificati per età, che forniscono lo stato immunologico corrente degli individui campionati. In particolare i valori ottimi dei parametri di trasmissione si ottengono massimizzando, condizionatamente alla matrice di contatto prescelta, la verosimiglianza dei dati sierologici calcolata a partire dalle prevalenze osservate di immuni per classe di età e dalle corrispondenti prevalenze attese, che sono funzioni del vettore dei parametri da stimare, del modello matematico.
Dalle stime dei parametri di trasmissione è possibile calcolare la corrispondente stima del tasso (o numero) di riproduzione di base dell'infezione, che rappresenta il numero medio di contatti adeguati che un infettivo ha nel suo periodo di infettività in una popolazione completamente suscettibile. Il numero di riproduzione di base è il più importante parametro di sintesi della trasmissibilità di un'infezione, fondamentale per la definizione dei programmi di controllo.
I pochi (vista la disponibilità solo molto recente dei dati sui contatti sociali) esempi disponibili di procedure inferenziali su modelli di trasmissione stimati su dati di contatto sono essenzialmente basati sul bootstrap.

1.4 Obiettivi del lavoro

Un primo obiettivo del lavoro è quello di valutare quali siano le tipologie di contatti sociali che meglio spiegano i dati di infezione (disponibili nella forma di dati sierologici), al fine di approfondire il ruolo dei diversi tipi di contatti nella diffusione di infezioni a contatto diretto e, in ultima analisi, di contribuire a migliorare le strategie di parametrizzazione dei modelli matematici.
Si sono pertanto adattati vari tipi di modelli di trasmissione di dati, e confrontate le prestazioni dei modelli utilizzati mediante criteri di bontà di adattamento.

Il secondo obiettivo del lavoro è quello di fornire una valutazione generale delle proprietà degli approcci bootstrap nella inferenza su coefficienti di trasmissione di infezioni, ed i relativi modelli.
Per quanto a nostra conoscenza l'utilizzo del bootstrap nella letteratura sui modelli epidemiologici, molto concentrata nell'area medica e di sanità pubblica, è di tipo meramente applicativo. In genere, le applicazioni si limitano ad utilizzare un determinato stimatore bootstrap, sovente lo stimatore più elementare, ovvero quello Normale, per determinare un intervallo di confidenza di una stima della trasmissione. In particolare non sono state minimamente indagate le caratteristiche dell'approccio bootstrap per questa specifica classe di problemi. Nella seconda parte della tesi pertanto si costruiscono diverse tipologie di intervalli di confidenza bootstrap pubblicati nella letteratura (normal Standard, studentized, Efron percentile, Bias corrected and accelerated), e li si confrontano al fine di fornire una valutazione della bontà relativa dei vari tipi di intervalli di confidenza nelle procedure inferenziali sul modello epidemiologico di base.
Inoltre si valutano i ruoli dei parametri del ricampionamento ed il ruolo mutuo delle due fonti di incertezza caratteristiche dei modelli epidemiologici (incertezza nei dati di contatto ed in quelli sierologici) sulla incertezza complessiva delle stime dei parametri di trasmissione.

1.5 Risultati

Per la varicella in Italia sono state ottenute nuove stime dei coefficienti di trasmissione per una varietà di modelli ed l'individuate le matrici dei contatti sociali che "spiegano" meglio i dati mediante criteri di model selection (AIC,BIC) & inferenza multi-model.
I modelli "migliori" sono il modello per prossimità del contatto (fisici/non fisici) e quello per luogo del contatto (casa,lavoro,scuola,altri luoghi). Non sembrano rilevanti i contatti di tipo non fisico e quelli nel tempo libero . Questo conferma congetture tradizionali per cui le infezioni dei bambini si trasmettono soprattutto a scuola ed a casa.
Le stime degli standard error bootstrap delle repliche dei parametri di trasmissione hanno confermato che per questi modelli la maggiore fonte di variabilità è dovuta ai contatti sociali.
Questo risultato può essere spiegato sia considerando la minore numerosità campionaria utilizzata nell'indagine Polymod (845 partecipanti italiani) rispetto all'indagine ESEN 2(2446 partecipanti), sia constatando che ciò che differenzia realmente un modello da un altro sono le tipologie di matrici dei contatti scelte,mentre i profili di seroprevalenza osservati sono gli stessi.
Analizzando il contributo marginale di ciascuna fonte si possono valutare e confrontare le velocità di convergenza corrispondenti ed investigare possibili relazioni di proporzionalità tra ordine di grandezza degli SE bootstrap marginali e velocità di convergenza .
In presenza di due fonti di incertezza sono stati calcolati una serie di statistiche bootstrap (SE, cv, bias, rapporto tra bias e SE) dei parametri di trasmissione capaci di aiutarci nella costruzione degli intervalli di confidenza e nella valutazione del numero minimo di ricampionamenti bootstrap necessari a giungere ai valori 'ideali' delle stime .

1.6 Ringraziamenti

Ringrazio il mio supervisore prof. Piero Manfredi del Dipartimento di Statistica e Matematica Applicata all'Economia dell'Università di Pisa per avermi seguito con attenzione fino dai tempi della laurea specialistica, per avermi introdotto ai dettagli tecnici della modellistica matematica delle infezioni, e per avermi consentito di utilizzare per questa tesi i suoi materiali per i corsi di perfezionamento nazionali ed internazionali. Ringrazio Emanuele del Fava per i molti suggerimenti nel mese passato full immersion a lavorare sulle stime dei modelli epidemiologici, Emilio Zagheni per le discussioni sul bootstrapping, e Giorgio Guzzetta per il supporto nella introduzione alla programmazione Matlab.

Ringrazio il Dipartimento di Statistica e Matematica Applicata all'Economia dell'Università di Pisa e Giuseppe Maccioni per avermi consentito l'accesso alla Sala di Calcolo.
Ringrazio il Dipartimento di Statistica 'G. Parenti' dell'Università di Firenze per avermi concesso l'opportunità di assistere a numerosi ed interessanti seminari e per aver potuto frequentare i Corsi di Perfezionamento organizzati per la formazione dei dottorandi.

2 La varicella

2.1 Sintomi, decorso clinico e complicanze

La varicella (www.WHO.org) è una malattia infettiva altamente contagiosa provocata dal virus Varicella zoster (VZV), della famiglia degli Herpes virus. Insieme a rosolia, morbillo, pertosse e parotite, la varicella è annoverata fra le malattie contagiose dell'infanzia, che nella maggioranza dei casi colpiscono i bambini tra i 5 e i 10 anni. L'uomo è l'unico serbatoio noto di questo virus: la malattia si trasmette quindi soltanto da uomo a uomo. Dopo un'incubazione di 2/3 settimane, la malattia esordisce con un esantema cutaneo (o rash, Fig. 2.1), febbre non elevata e sintomi generali solitamente lievi, come malessere e mal di testa. Per 3-4 giorni, piccole papule rosa pruriginose compaiono su testa, tronco, viso e arti, ad ondate successive. Le papule evolvono in vescicole, poi in pustole e infine in croste granulari, destinate a cadere. Tipicamente l'esantema è costituito da 250-500 lesioni.

Fig. 2.1 :Esantema Cutaneo

La varicella è in genere una malattia benigna che guarisce nel giro di 7-10 giorni. La malattia tende ad avere un decorso più aggressivo al crescere dell'età, e quindi nell'adolescente e nell'adulto rispetto ai bambini più piccoli. La malattia può essere particolarmente grave in soggetti immunodepressi (ad esempio soggetti con infezione da Hiv/AIDS, soggetti sottoposti a trattamento chemioterapico, o in cura con steroidi per asma o altre malattie). Le complicanze della varicella tuttavia sono rare nei bambini sani e si verificano per lo più nelle persone immunodepresse, nei neonati e negli adolescenti o adulti. Tra le complicazioni riportate nella letteratura medica possono verificarsi superinfezione batterica delle lesioni cutanee, trombocitopenia, artrite, epatite, atassia cerebellare, encefalite, polmonite e glomerulonefrite. Tra gli adulti la complicanza più comune è la polmonite.
L'infezione produce immunità permanente in quasi tutte le persone immunocompetenti: raramente una persona può sviluppare due volte questa malattia. Tuttavia, il virus non viene eliminato dall'organismo, ma rimane latente (in genere per tutta la vita) nei gangli delle radici nervose spinali. Nel 10-20% dei casi il virus si riattiva a distanza di anni o di decenni, solitamente dopo i 50 anni, dando luogo al cosiddetto herpes zoster, noto comunemente come "fuoco di Sant'Antonio". L'herpes zoster si manifesta con lesioni a grappolo di tipo vescicolare al torace, a volte accompagnate da dolore localizzato. Il dolore che persiste oltre un mese viene chiamato neuralgia posterpetica.
Se la varicella viene contratta da una donna all'inizio di una gravidanza (fondamentalmente nel primo trimestre di gestazione) può trasmettersi al feto, causando una embriopatia nota come sindrome della varicella congenita. I bambini che sono stati esposti al virus della varicella in utero dopo la ventesima settimana di gestazione possono sviluppare una varicella asintomatica e successivamente herpes zoster nei primi anni di vita. Se invece la madre ha avuto la malattia da cinque giorni prima a due giorni dopo il parto, può verificarsi una forma grave di varicella del neonato, la cui mortalità può arrivare fino al 30%.

La varicella è una delle malattie infettive più contagiose. La contagiosità si verifica nei 4-5 giorni precedenti la comparsa dell'esantema (quando l'individuo è quindi ancora asintomatico) e poi nei primi stadi dell'eruzione. La trasmissione da persona a persona avviene per via aerea mediante le goccioline respiratorie diffuse nell'aria quando una persona affetta per esempio tossisce o sternutisce emettendo dell' articolato (Fig. 2.2), o tramite contatto diretto con lesione da varicella o zoster. Durante la gravidanza, il virus può essere trasmesso all'embrione o al feto attraverso la placenta.

Fig. 2.2:Fotografia chamber della traiettoria di uno sternuto

Madri immuni che hanno sperimentato l'infezione in qualunque fase della loro vita precedente alla gravidanza trasmettono i propri anticorpi al figlio attraverso la circolazione transplacentare. Questi anticorpi acquisiti generano una fase temporanea di immunità del neonato di durata circa 6 mesi.

2.2 Terapia e prevenzione

Essendo la varicella una infezione virale la terapia è solo sintomatica. Per il prurito possono essere utilizzati antistaminici, mentre per la febbre il paracetamolo. I bambini con varicella non devono essere trattati con salicilati (aspirina), perché questo aumenta il rischio di sindrome di Reye.
Nei casi più a rischio di complicanze (adolescenti, persone con malattie respiratorie croniche o in trattamento con steroidi) e nei casi secondari familiari si può ricorrere a farmaci antivirali come l'acyclovir. La terapia antivirale non è raccomandata nei bambini con varicella altrimenti sani, visto che, somministrata per via orale entro 24 ore dall'inizio dell'esantema, determina solamente una modesta riduzione dei sintomi. Nei pazienti immunodepressi è raccomandata la terapia antivirale per via venosa.
In generale, si consiglia di isolare i pazienti per evitare la diffusione del contagio. È raccomandato che i bambini colpiti dalla malattia restino a casa da scuola per almeno cinque giorni dalla comparsa delle prime vescicole.
Dal 1995 è disponibile un vaccino, costituito da virus vivo attenuato che alcuni Paesi, tra cui gli Usa, raccomandano per tutti i bambini nel secondo anno di vita. L'efficacia della vaccinazione è stata stimata essere del 95%, nella prevenzione delle forme moderate o gravi; del 70-85% nella prevenzione delle forme lievi. Il vaccino è sicuro e ben tollerato e la protezione sembra essere di lunga durata. La vaccinazione va effettuata con una sola dose ai bambini tra 12 mesi e 12 anni, e con due dosi in chi ha più di 12 anni. Il vaccino è controindicato per gli individui immunodepressi, mentre è consigliato nei bambini più grandi, negli adolescenti e negli adulti che non abbiano ancora contratto la malattia e privi di controindicazioni. È consigliato soprattutto per le persone che per motivi professionali hanno un maggior rischio di acquisire l'infezione (come il personale scolastico) o trasmetterla a persone ad alto rischio di complicanze gravi (come gli operatori sanitari). Inoltre, la vaccinazione è particolarmente indicata anche per le donne in età fertile che non hanno già avuto la malattia, per evitare un'eventuale infezione in gravidanza e i conseguenti danni al bambino.
Per persone a elevato rischio di varicella grave (alcuni neonati, soggetti immunocompromessi) è raccomandato l'utilizzo di immunoglobine per via intramuscolare (immunoprofilassi passiva) se esposte a persone con la varicella. Queste vanno somministrate quanto prima e fino a 96 ore dopo l'esposizione.

La vaccinazione dei bambini suscettibili entro 72 ore e non oltre le 120 ore dall'esposizione può prevenire e modificare significativamente la malattia. L'acyclovir per via orale non è raccomandato come profilassi.

2.3 La diffusione della varicella e le strategie di prevenzione e vaccinazione

La varicella è molto diffusa in tutto il mondo temperato. In Italia si verificano epidemie annuali, con incidenza massima in primavera. Il 90% dei casi notificati riguarda bambini e ragazzi fino a 14 anni. I risultati del sistema di sorveglianza sentinella Spes (www.salute.gov) mostrano infatti che ogni anno la varicella interessa il 5% circa della popolazione di questa fascia di età. Questo corrisponde a una incidenza media di circa 500.000 casi per anno, in accordo con i risultati di studi relativi alla notifica dell'infezione, e con studi basati su dati di sieroprevalenza nella popolazione generale e su modelli matematici. La fascia di età più colpita è quella tra 1 e 4 anni. Anche se più rara, la varicella può colpire anche ragazzi più grandi e adulti; in particolare, i dati di sieroprevalenza mostrano che circa il 10% della popolazione tra 20 e 40 anni non ha ancora contratto l'infezione, ed è quindi a rischio di ammalarsi. In età pediatrica, la varicella è una malattia relativamente benigna; la frequenza di complicanze stimata da uno studio italiano condotto negli anni '90 è infatti pari al 3,5%, mentre quella dei ricoveri è dello 0,9%. Vi sono inoltre studi internazionali (Marta Ciofi, 2008) che mostrano nei bambini una frequenza di complicanze severe e di decessi rispettivamente di 8 e 2 casi ogni 100.000 malati (pari cioè allo 0,008% e 0,002%). La gravità della malattia aumenta invece con l'età, e negli adulti la frequenza di complicanze, ricoveri e decessi è stimata essere rispettivamente 7, 9 e 25 volte superiore rispetto ai bambini. Inoltre, la varicella può avere un decorso particolarmente grave nelle persone immunodepresse di qualsiasi età. L'infezione può essere prevenuta con il vaccino costituito da virus vivo attenuato, che in alcuni Paesi, tra cui gli Usa, è raccomandato per tutti i bambini nel secondo anno di vita. L'efficacia della vaccinazione nei bambini è stata stimata essere del 93% in un trial clinico controllato, e del 73% circa in studi di campo. La vaccinazione, che come gli altri vaccini vivi attenuati è controindicata negli individui con deficit della risposta immune, va effettuata con una sola dose ai bambini tra 12 mesi e 12 anni, e con due dosi in chi ha più di 12 anni. Modelli matematici internazionali e nazionali mostrano che la vaccinazione su larga scala per i nuovi nati andrebbe attuata solo se si possono raggiungere in tempi brevi coperture vaccinali superiori all'80% in ogni coorte di nascita. In caso contrario si verificherebbero effetti indesiderati, quali lo spostamento in avanti dell'età dei casi, con una maggiore incidenza in età in cui la malattia è più grave. La vaccinazione degli adolescenti, pur avendo un impatto modesto sull'incidenza totale della malattia, consente invece di ridurre la frequenza dei casi a maggior rischio di complicanze. Considerata la maggior gravità della malattia all'aumentare dell'età, la vaccinazione degli adulti suscettibili rappresenta un'azione prioritaria, soprattutto per le persone che per motivi professionali hanno un maggior rischio di acquisire l'infezione (come il personale scolastico) o trasmetterla a persone in fragili condizioni di salute (come gli operatori sanitari). Anche le donne in età fertile rappresentano un gruppo di popolazione per cui la vaccinazione è particolarmente necessaria, perché l'infezione in gravidanza può trasmettersi al feto causando una embriopatia, se la varicella è stata acquisita nei primi due trimestri di gestazione, o una forma grave di varicella del neonato se la madre ha avuto la malattia da 5 giorni prima a due giorni dopo il parto. In questo caso, la mortalità del neonato può arrivare fino al 30%. Inoltre, analogamente a quanto accade per altre malattie infettive, come il morbillo o le meningiti meningococciche, è presumibile che anche chi vive in comunità chiuse (caserme, carceri, collegi universitari ecc) abbia un maggior rischio di contrarre l'infezione. A questo proposito bisogna ricordare che, come il vaccino contro il morbillo, anche quello contro la varicella è efficace nella profilassi post-esposizione, se somministrato entro 3 giorni. Se si verifica un caso in una comunità chiusa di adulti, è quindi opportuno vaccinare chi non ricorda di avere avuto la varicella. L'anamnesi di mancata malattia, infatti, è affidabile per identificare i suscettibili.

3 I dati sierologici

3.1 I dati di seroprevalenza

Il tipo di dato più utilizzato in epidemiologia delle malattie infettive per documentare la circolazione di un'infezione con immunità permanente è il cosiddetto dato sierologico. Il dato sierologico è un dato di tipo "current-status" che fornisce lo stato immunologico corrente dell'individuo campionato L'obiettivo di un esame sierologico è quello di accertare la presenza di anticorpi prodotti dall'organismo in risposta agli specifici antigeni responsabili dell'infezione. L'organismo può produrre molti tipi di anticorpi, tra cui quelli di tipo IgG(ImmunogammaGlobuline), che vengono prodotti dopo che l'individuo è stato infettato. Nelle infezioni che generano immunità permanente i titoli anticorpali tendono ad assumere inizialmente valori molto alti, e successivamente, pur decadendo numericamente nel tempo, rimangono rilevabili per un periodo di tempo molto lungo (anche se non è chiaro se e in quale misura vengano aumentati nel corso di successive esposizioni dell'individuo immune con soggetti infettivi). Nelle analisi sierologiche il ricercatore misura il livello di anticorpi nel sangue di un individuo. Se il singolo individuo presenta un "alto" (in base a criteri prestabiliti, variabili da infezione a infezione) livello di anticorpi (ovvero sopra una soglia assegnata A_2) contro una specifica malattia, è classificato come immune (ossia si assume che abbia sperimentato l'infezione nel passato se in regime pre-vaccinale, o che sia stato vaccinato contro l'infezione). Al contrario se la quantità di anticorpi è sotto un quantità soglia specifica A_1, allora l'individuo è classificato come "non sufficientemente protetto", e quindi come "suscettibile". Infine individui con titolazione anticorpale compresa tra A_1 ed A_2 sono classificati come "indecisi" (principio del cut-off).
Mediante analisi delle titolazioni anticorpali di un campione rappresentativo di individui è quindi possibile documentare il grado, ossia tecnicamente, *la prevalenza, di immunità* nella popolazione.
Esistono 2 tipi di indagini sierologiche : studi di tipo longitudinale ('**follow up**') e di tipo trasversale ('**cross sectional**').
Negli studi '**follow up**' il ricercatore seleziona un campione di individui suscettibili (soggetti che non presentano anticorpi contro la malattia) che provengono dalla stessa coorte (d'età) e quindi segue tali individui per un periodo di tempo. Durante questo periodo, il ricercatore registra quando un individuo viene infettato e quindi il suo organismo inizia a produrre anticorpi o registra lo stato di suscettibilità del soggetto quando il livello di anticorpi è pressoché nullo: questo prende il nome di periodo a rischio, perché durante questo periodo il soggetto rischia di acquisire l'infezione. Alla fine si conosce chi nel campione è sieropositivo (persone che presentano un livello alto di anticorpi) o no, e l'età dell'infezione per gli individui sieropositivi. In questo modo il ricercatore può valutare la percentuale di persone con anticorpi ad ogni età. Tramite un' indagine follow-up è possibile stimare due parametri fondamentali di grande interesse per gli epidemiologi: la sieroprevalenza ,che è una frequenza e misura la percentuale di persone con anticorpi in un specifico momento nella popolazione ; la forza d'infezione che è un tasso di rischio (o tasso d'incidenza) e misura il numero di nuovi infetti durante un certo periodo di tempo nella popolazione campionaria. Lo studio 'follow up' ha il vantaggio di ridurre l'errore sistematico, perché il ricercatore può controllare personalmente la qualità dei dati durante l'indagine; il grande svantaggio è invece la quantità di tempo necessario per concludere l'indagine, affinché si registri l'avvenuta infezione.
Negli studi sierologici trasversali, il ricercatore seleziona un campione, tipicamente stratificato per età, di individui in un istante assegnato del tempo, e valuta come gli individui campionati sono distribuiti (per ogni classe di età) tra sieropositivi (immuni) e sieronegativi (suscettibili). Gli studi trasversali hanno il vantaggio di richiedere molto meno tempo rispetto ad uno studio follow up, giacchè l'indagine è interessata unicamente allo "stato corrente" dei soggetti campionati. Per questo motivo gli studi cross-sectional sono molto più comuni rispetto agli studi follow up.
In particolare se si vuole stimare la forza di infezione (ovvero l'hazard di infezione) in presenza di dati ottenuti da uno studio trasversale, occorre naturalmente assumere che i dati siano rappresentativi dei cambiamenti longitudinali nella sieroprevalenza con l'età. Infatti il campione

sierologico trasversale descrive tecnicamente quella che in demografia chiameremmo una coorte "fittizia", ottenuta dall'accostamento di molte coorti di nascita distinte. Pertanto solo assumendo la omogeneità dei comportamenti delle varie coorti possiamo utilizzare il dato trasversale per inferire il corrispondente sottostante comportamento longitudinale.
Formalmente una indagine sierologia modella ogni soggetto campionato mediante una variabile dicotomica (1= individuo è immune, 0=l'individuo è suscettibile), valendo quanto detto sopra per i soggetti "indecisi". Tramite la dicotomizzazione,è possibile stimare la prevalenza di immuni per ognuna delle classi d'età considerate. Torneremo su questi aspetti più tecnici nel corso dei capitoli successivi.

3.2 I dati italiani di sierologia della varicella

I dati italiani utilizzati in questo lavoro provengono dall'indagine ESEN 2 (2004), un progetto finanziato dall'Unione Europea nell'ambito del Quinto Programma Quadro. L'ampiezza del campione sierologico italiano è di 2446 unità, stratificato per classi d'età annuali. La Fig. 3.1 riporta il diagramma a dispersione della prevalenza di soggetti immuni alla varicella al variare dell'età nel campione italiano dell'indagine ESEN2. Come si può notare la proporzione di immuni nel campione tende mediamente a crescere al crescere dell'età, a documentare la natura cumulatica del processo di acquisizione di immunità al trascorrere dell'età. Tuttavia la velocità di crescita del profilo sierologico osservato appare molto differenziata al variare dell'età, con una crescita molto veloce nei primi anni di vita (la soglia cumulativa del 90% è raggiunta poco sopra i 10 anni di vita) e poi progressivamente sempre più lenta alle età adulte ed anziane. Oltre i 30 anni il grafico mostra un grande rumore che impedisce di valutare se effettivamente il processo di infezione con immunità sia completo (=probabilità cumulativa di infezione del 100%) o difettivo.

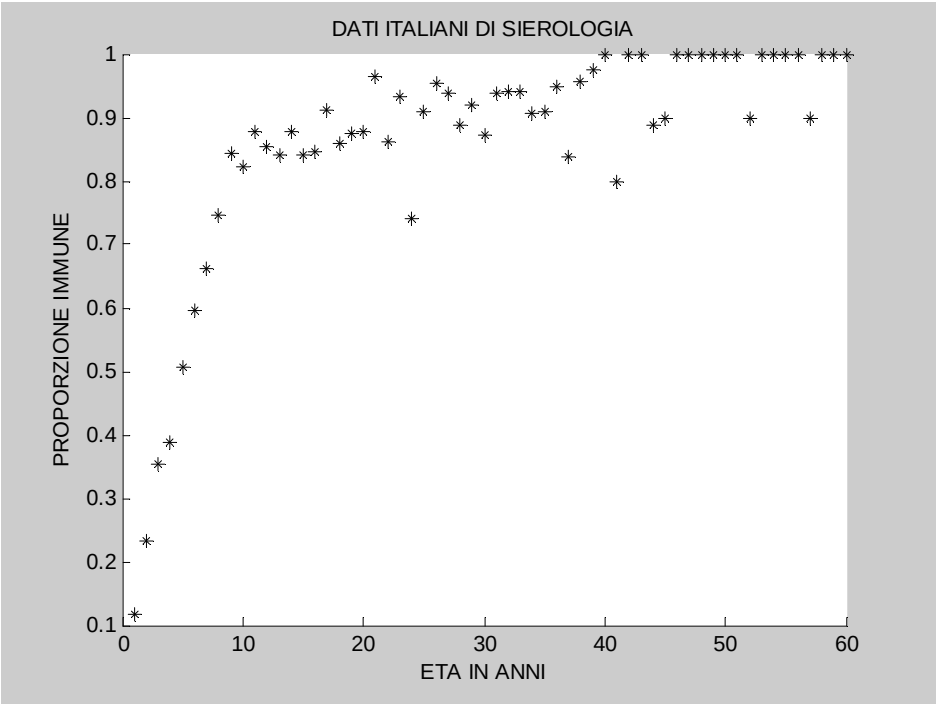


Fig. 3.1: Varicella in Italia. Proporzioni di soggetti immuni per classi d'età annuali.

La Fig. 3.2 mostra le numerosità campionarie (espresse in frequenze relative) delle singole classi di età annuali del campione sierologico italiano dell'indagine ESEN 2. Il grafico di Fig. 3.2 mostra una tipica prassi delle indagini sierologiche di sovracampionare gli individui di età inferiore a 10 anni, essenzialmente motivata dall'interesse medico per la varicella in quanto malattia dell'infanzia. Il sottocampionamento delle classi di età anziane è responsabile del forte rumore descritto in Fig. 3.1.

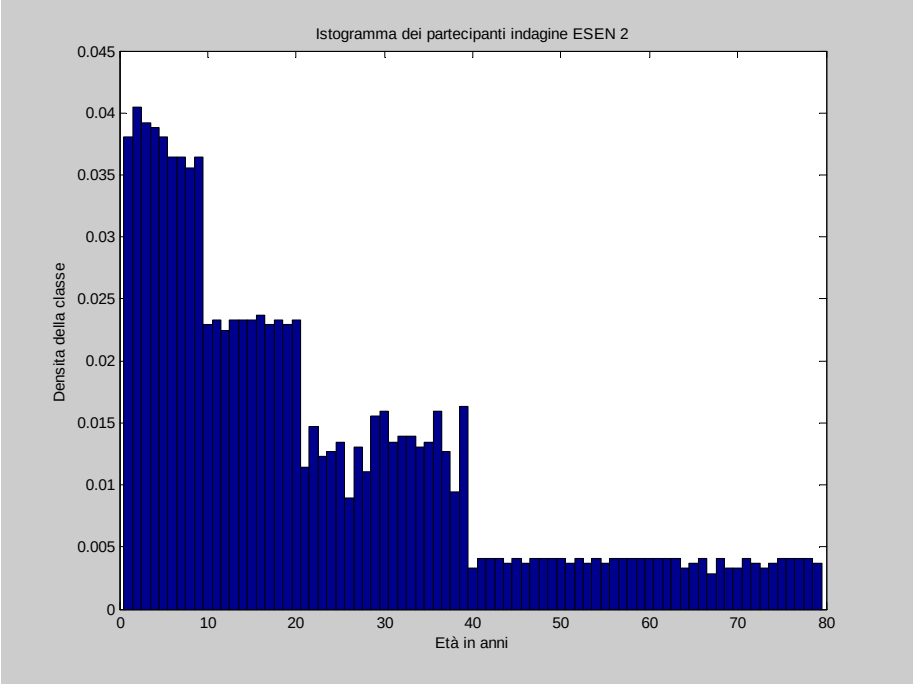


Fig. 3.2 :Istogramma dei partecipanti per classe ESEN2

Nelle analisi successive assumeremo, per i fini delle nostre analisi, che il campione sierologico sia statisticamente "well-behaved", ossia costituito da osservazioni "essenzialmente" indipendenti ed identicamente distribuite. Tuttavia è bene ricordare che motivi di costo rendono proibitive analisi sierologiche di questo tipo. Per esempio i dati della maggior parte dei paesi dell'indagine ESEN 2 sono dati sierologici del tipo cosiddetto "residuale", ovvero provenienti da analisi di sieri disponibili nelle banche ospedaliere, tipicamente selezionati senza criteri di casualità del campionamento.

4 Il modello matematico delle infezioni virali infantili

4.1 Introduzione

In assenza di vaccinazione, quando era possibile ipotizzare che essenzialmente tutti gli individui della popolazione sperimentassero la varicella nel corso della vita, il corso individuale dell'infezione era caratterizzato dalla seguenti fasi: (a) la fase di **protezione anticorpale materna (M)**, immediatamente successiva alla nascita in cui fondamentalmente tutti gli individui erano protetti dagli anticorpi materni acquisiti a seguito della trasmissione transplacentare materno-fetale (nell'ipotesi che la madre avesse acquisito l'infezione con probabilità 1); dalla fase M, la cui durata è tipicamente di pochi mesi (in dipendenza del tempo medio di decadimento della dotazione anticorpale), quale l'individuo entra nella fase di **suscettibilità (S)**, nel corso della quale è suscettibile all'acquisizione dell'infezione come conseguenza di contatti sociali del tipo appropriato con soggetti infettivi. Individui suscettibili che hanno acquisito l'agente infettivo avuto dei contatti entrano in una fase (durata: 5-7 giorni) di **latenza non infettiva (E)**, durante la quali si osserva la rapida replicazione virale all'interno dell'individuo, a seguito della quale l'individuo diventa **infettivo (I)** a tutti gli effetti e quindi in grado di ritrasmettere l'infezione ad altri. La maggior parte della fase infettiva (durata circa 7 giorni) è *asintomatica*, il che impedisce l'isolamento dell'individuo e consente quindi il dispiegamento della gran parte del suo potenziale infettivo. Solo nell'ultima parte della fase infettiva appaiono i sintomi caratteristici della *fase di malattia* (da varicella), nella forma di esantema, che tipicamente portano all'isolamento dell'individuo, che ne riduce i contatti a rischio (per esempio se l'isolamento è nella casa di abitazione contatti a rischio saranno solo con i famigliari, dopo però svariati giorni di contatti nei giorni precedenti). L'infettività si riduce a zero già prima della conclusione dell'esantema, la cui **rimozione dal circuito infettivo (R)** è identificata con la guarigione dell'individuo, a seguito della quale l'individuo acquisisce *immunità permanente*. La rimozione rappresenta lo stato assorbente non banale del sistema. Il diagramma a blocchi, o "compartimenti" riportato in Fig. 4.1, che sintetizza la sequenza dei passaggi di stato caratteristici delle infezioni esantematiche infantili ad immunità permanente, costituisce la base delle descrizioni matematiche delle infezioni.

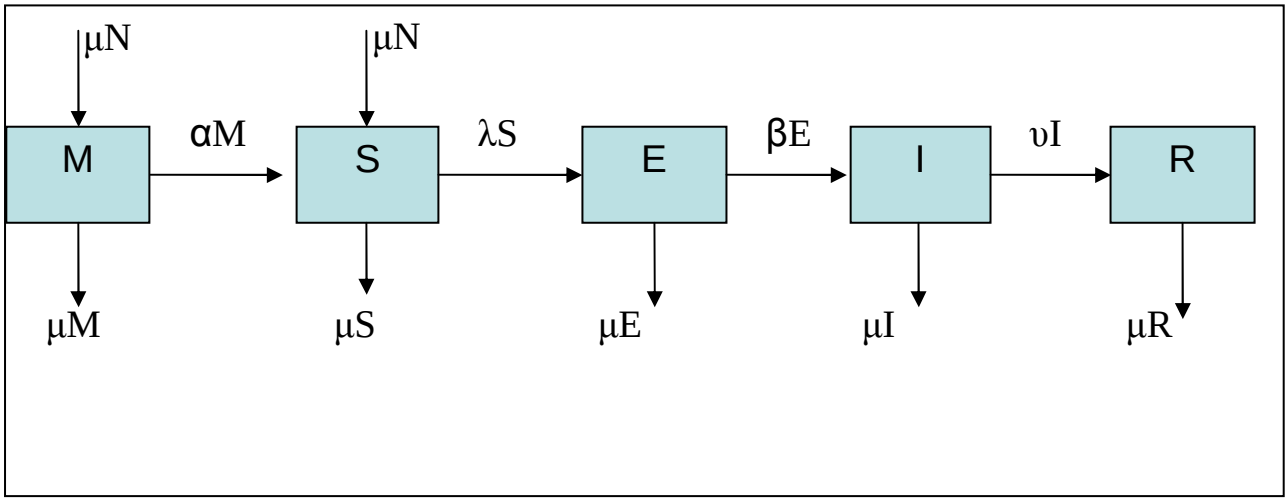


Fig. 4.1 :Sequenza MSEIR di passaggi di stato tipici delle infezioni esantematiche infantili in assenza di vaccinazione.

La rappresentazione MSEIR è particolarmente appropriata per le infezioni virali infantili, grazie allo scandimento molto marcato delle varie fasi di infezione (più complesse per altri tipi di infezioni).

Tuttavia per scopi pratici, in particolare per problemi di stima dei parametri, conviene utilizzare una rappresentazione semplificata. Conviene in particolare trascurare le fasi M ed E, in quanto fondamentalmente irrilevanti. Per la fase E questo è del tutto ovvio, in quanto non osservabile da osservazioni di popolazione, e quindi incorporabile nella fase S. Per la fase M possiamo semplicemente assumere che il momento della nascita coincida con il momento in cui mediamente si verifica la perdita degli anticorpi materni, e quindi con il momento di ingresso in suscettibilità. In questo modo otteniamo la classica rappresentazione Suscettibile-Infettivo-Rimosso (SIR), il cui diagramma a blocchi è riportato in Fig. 4.2,con i corrispondenti tassi di transizione di stato.

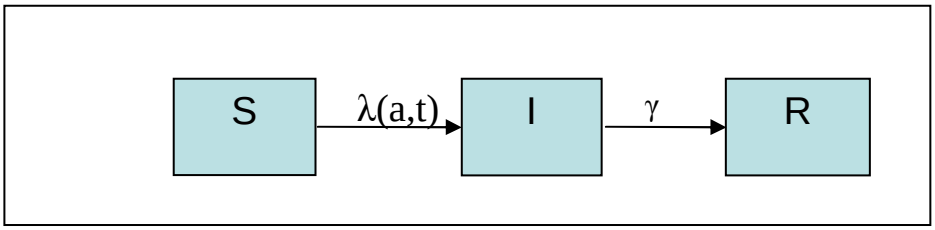


Fig. 4.2Diagramma a compartimenti del modello SIR con i corrispondenti tassi di transizione di stato: la forza dell'infezione, e la forza di rimozione.

In particolare il tasso di transizione $\lambda(a,t)$ dal comparto suscettibile al comparto infettivo, che costituisce quindi la "probabilità istantanea" di passaggio di stato (quindi tecnicamente un hazard rate), è tradizionalmente chiamato la forza dell'infezione (FOI). La FOI è tipicamente assunta funzione sia dell'età sia del tempo. La dipendenza dall'età è stata documentata in numerosi studi sulle infezioni infantili (Anderson e May 1991, Edmunds et al., 2000) in conseguenza, in particolare, dell'elevato numero di contatti "adeguati" alla trasmissione nel corso delle fasi di frequenza scolare. La dipendenza dal tempo è primariamente la conseguenza dell'andamento non costante delle infezioni nella popolazione, in conseguenza del carattere epidemico ricorrente, con periodicità precise, delle infezioni virali infantili. Come noto soprattutto dagli studi sul morbillo le dinamiche "*naturali*" (ossia in assenza di controlli umani, quali le vaccinazioni) delle infezioni in popolazioni sufficientemente grandi sono caratterizzate dal susseguirsi di fasi *epidemiche* di breve durata (pochi mesi) caratterizzate da una rapida crescita del numero di infettivi e dei casi osservati sintomatici di malattia, cui seguono delle fasi di recupero caratterizzate dal mantenimento dell'infezione a livelli molto bassi. Altri fattori time-dependent che possono influenzare in maniera più strutturale la forza dell'infezione sono naturalmente i vaccini, i cui effetti peraltro non verranno presi in considerazione in questo lavoro. Torneremo tra poco sulla complessa dipendenza che lega la FOI con le dinamiche degli stati infettivi. Il tasso γ. di transizione dalla fase infettiva alla fase immune, detto anche tasso di guarigione, è invece tipicamente assunto costante, come conseguenza di una distribuzione di tipo esponenziale della durata della fase infettiva. Pertanto vale la relazione $\gamma=1/D$, ove D è la durata attesa del periodo infettivo. Empiricamente D=7 giorni, cosi che $\gamma=(1/7)\text{giorno}^{-1}=52\text{anno}^{-1}$. Il principale tratto caratteristico delle infezioni virali infantili con immunità permanente è dunque che una brevissima fase infettiva (7 giorni per varicella, morbillo, etc.) separa due fasi invece molto lunghe della vita dell'individuo, ossia la fase iniziale di suscettibilità, tipicamente di alcuni anni (l'età media all'infezione della varicella nel mondo industrializzato è dell'ordine di 6-8 anni) e la fase finale di immunità. In questa prospettiva un processo di infezione con immunità permanente risulta quindi a tutti gli effetti un processo di eliminazione di una coorte di sopravviventi (i suscettibili) per acquisizione di immunità, e la forza di infezione il corretto (salvo una piccolissima approssimazione dell'ordine del rapporto tra la durata della fase infettiva e la durata della vita) tasso di transizione dalla fase suscettibile a quella immune. Questa prospettiva del tipo "Suscettibile-Rimosso" (SR) (quindi una catena di Markov omogenea con uno stato transiente ed uno assorbente) può essere adottata per il mero scopo della stima della forza di

infezione. Tuttavia questa prospettiva della sopravvivenza non è come vedremo particolarmente utile dal punto di vista della modellazione delle infezioni. La ragione principale è che la stima della sola FOI non informa adeguatamente sulla struttura del sottostante modello matematico di trasmissione dell'infezione, dove la classe trascurata, quella degli infettivi, gioca un ruolo centrale, che differenzia un processo di infezione da un puro processo di sopravvivenza. Questo impedisce le principali utilizzazioni del modello matematico, ossia per esempio la stima del Numero di riproduzione di base R_0 (Diekmann e Heesteerbeek, 1990), e la valutazione dell'impatto di un programma di vaccinazione (Anderson e May, 1991). Come vedremo per questi scopi occorre stimare l'intero modello di trasmissione inteso nel senso della relazione tra FOI e le dinamiche degli stati infettivi. Nelle sezioni seguenti introduciamo pertanto i dettagli del modello matematico di trasmissione.

4.2 Il modello SIR per le dinamiche naturali pre-vaccinali dell'infezione

Indichiamo con *X(a,t),Y(a,t),Z(a,t)* le funzioni del tempo e dell'età che descrivono l'andamento nel tempo del numero (per l'esattezza si tratta di una densità assoluta) di soggetti di età esatta a che sono rispettivamente suscettibili, infettivi e rimossi al tempo t. Le transizioni epidemiologiche spiegate nella sezione precedente sono descritte dalle seguenti equazioni alle derivate parziali con condizioni al contorno (Anderson e May, 1991)

$$\begin{cases} \left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)X(a,t) = -(\mu(a)+\lambda(a,t))X(a,t) & , X(0,t)=B \\ \left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)Y(a,t) = \lambda(a,t)X(a,t)-(\mu(a)+\gamma)Y(a,t) & , Y(0,t)=0 \\ \left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)Z(a,t) = \gamma Y(a,t)-\mu(a)Z(a,t) & , Y(0,t) \end{cases} \quad (4.1)$$

dove il significato dei tassi di transizione γ,λ è già stato spiegato; in aggiunta *μ(a)* rappresenta il tasso di mortalità specifico dell'età *a*, *n(a)* la densità assoluta di popolazione di età a, e infine B il numero di ingressi per natalità nella popolazione al tempo t. Le condizioni al contorno:

$$\begin{cases} X(0,t)=B \\ Y(0,t)=0 \\ Z(0,t)=0 \end{cases} \quad (4.2)$$

sono le tipiche condizioni al bordo delle infezioni virali con immunità (ed assenza di trasmissione verticale dell'infezione), nel loro regime naturale "pre-vaccinale": per definizione tutti gli individui che entrano nella popolazione entrano nel comparto suscettibile.
In particolare sommando le tre equazioni troviamo l'equazione che esprime la dinamica della popolazione complessiva $n(a,t)=X(a,t)+Y(a,t)+Z(a,t)$:

$$\left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)n(a,t) = -\mu(a)n(a,t) \quad , \quad n(0,t)=B \quad (4.3)$$

Se il numero di nascite per unità di tempo è costante tale equazione descrive una popolazione stazionaria, come assumeremo definitivamente d'ora in poi, con struttura di età assoluta:

$$n(a) = Be^{-\int_0^a \mu(x)dx} \qquad (4.4)$$

Infine la composizione relativa di età è la nota:

$$c(a) = \frac{n(a)}{\int_0^\omega nada} = \frac{Be^{-\int_0^a \mu(x)dx}}{\int_0^\omega \left(Be^{-\int_0^a \mu(x)dx} \right) da} = \frac{p(a)}{e_0} \qquad (4.5)$$

Dove ω, *p(a)* ed e_0 rappresentano rispettivamente l'età massimale, la funzione di sopravvivenza e la speranza di vita alla nascita. In particolare la stazionarietà della popolazione consente di semplificare il sistema eliminando lo stato assorbente R mediante la identità:

$$Z(a,t) = n(a,t) - X(a,t) - Y(a,t) \qquad (4.6)$$

Se in particolare assumiamo che la mortalità sia del cosiddetto tipo 1, ovvero caratterizzata da una funzione di sopravvivenza costante all'unità fino all'età massimale (Anderson e May, 1991), che costituisce un modello di mortalità grossolano ma molto appropriato per la descrizione delle dinamiche di trasmissione in paesi industrializzati in assenza di vaccinazione, quando essenzialmente tutta la trasmissione si manifestava prima dei 15 anni e quindi molto prima che la mortalità iniziasse a crescere significativamente, il modello assume la forma semplificata

$$\begin{cases} \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t} \right) X(a,t) = -\lambda(a,t) X(a,t) & , X(0,t) = B \\ \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t} \right) Y(a,t) = \lambda(a,t) X(a,t) - \gamma Y(a,t) & , Y(0,t) = 0 \\ \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t} \right) Z(a,t) = \gamma Y(a,t) & , Y(0,t) \end{cases} \qquad (4.7)$$

La forma (4.7) precedente è particolarmente conveniente per le analisi empiriche.
Alternativamente il modello può essere espresso nelle frazioni epidemiologiche: *S(a,t)=X(a,t)/n(a), I(a,t)=Y(a,t)/n(a), R(a,t)* esprimente le frazioni di individui suscettibili, infettivi e rimossi di età *a* presenti nella popolazione al tempo t. Con semplici passaggi si trova che la struttura del modello nelle frazioni epidemiologiche è la seguente:

$$\begin{cases} \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t} \right) S(a,t) = -\lambda(a,t) S(a,t) & , S(0,t) = 1 \\ \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t} \right) I(a,t) = \lambda(a,t) S(a,t) - \gamma I(a,t) & , I(0,t) = 0 \\ R(a,t) = 1 - S(a,t) - I(a,t) & \end{cases} \qquad (4.8)$$

Dove le condizioni al contorno si modificano come:

$$\begin{cases} S(0,t)=1 \\ I(0,t)=0 \\ R(0,t)=0 \end{cases} \qquad (4.9)$$

Per completare lo formulazione del modello dinamico, nei numeri oppure nelle frazioni, sono necessarie appropriate condizioni iniziali, che specifichino le distribuzioni di età di suscettibili infetti e rimossi ad un assegnato tempo zero iniziale.
Il parametro centrale del modello, la forza di infezione $\lambda(a,t)$, è essa stessa una variabile di stato del sistema. La più comune definizione comportamentale della FOI è la seguente:

$$\lambda(a,t)=\int_0^{\infty} q(a,a')C(a,a')I(a',t)da' \,. \qquad (4.10)$$

ove $C(a,a')$ è il *numero medio di contatti sociali* intrattenuti per unità di tempo da un soggetto di età *a* con soggetti di età *a'*, mentre il coefficiente $q(a,a')$ è il rischio *infection-specific di trasmissione* dell'infezione nel corso di un singolo contatto tra soggetti del tipo (a,a'). La interpretazione della precedente espressione della FOI è che il rischio atteso di acquisire l'infezione (per unità di tempo) da parte di un soggetto suscettibile di età *a* nel corso di contatti sociali con soggetti di età a' dipende dal prodotto del numero di *contatti a rischio*, ossia con soggetti infettivi, che si trova moltiplicando il numero di $C(a,a')$ per la frazione infettiva $I(a',t)$, moltiplicato a sua volta per la probabilità $q(a,a')$ che un singolo contatto a rischio generi effettivamente la trasmissione.
Poiché anche i migliori dati disponibili sui contatti sociali sono disponibili su classi di età piuttosto grossolane, ma anche per certi motivi tecnici relativi alla trattazione matematica del modello e che diverranno evidenti più tardi, si preferisce tipicamente considerare in luogo della precedente espressione la sua discretizzazione per classi di età di ampiezza finita:

$$\lambda_i(t)=\sum_{j=1}^{m} q_{ij}C_{ij}I_j(t), \qquad (4.11)$$

dove $[a_{i-1},a_i)$ è la generica classe di età (con ampiezza $h_i=a_i-a_{i-1}$), C_{ij} il corrispondente numero medio di contatti tra soggetti di età i e j (descritti a questo punto da una opportuna *matrice dei contatti*, q_{ij} i corrispondenti coefficienti di trasmissione specifici per età, ed $I_j(t)$ la frazione di soggetti infettivi nella classe di età j-esima.
Non siamo in questa sede interessati alle proprietà dinamiche del modello SIR con struttura di età, che si studiano preferibilmente con tecniche di analisi matematica avanzata (Inaba, 1990). Nella seguente sezione ne sintetizziamo però alcuni tratti fondamentali con la discussione dei suoi stati di equilibrio. L'analisi di equilibrio è quella sui cui si fondano le analisi statistiche dei dati di trasmissione.

4.3 Il modello SIR: dinamiche di equilibrio

La dinamica di equilibrio del modello si ottiene eliminando le dipendenze temporali (quindi in particolare ponendo uguali a zero le derivate parziali rispetto al tempo). Ne risulta quindi un sistema di 2 equazioni differenziali ordinarie a coefficienti non costanti e valori iniziali. Lavorando per esempio sul modello nelle frazioni (che peraltro è equivalente, salvo una costante di scala) al modello con mortalità del tipo 1:

$$\begin{cases} S'(a) = -\lambda(a)S(a) & S(0)=1 \\ I'(a) = \lambda(a)S(a) - \gamma I(a) & I(0)=0 \\ R(a) = 1 - S(a) - I(a) \end{cases} \qquad (4.12)$$

Come si vede facilmente il modello possiede sempre uno stato di equilibrio di eliminazione dell'infezione ("disease free equilibrium" o DFE) $E_0 = (1,0,0)$ in cui le frazioni immuni ed infettive sono sempre uguali a zero. Questo stato di equilibrio genera una forza dell'infezione identicamente uguale a zero, come atteso. Inoltre il modello può possedere altri stati di equilibrio che possono essere trovati risolvendo formalmente il modello:

$$\begin{aligned} S(a) &= e^{-\int_0^a \lambda(u)du} \\ I(a) &= \int_0^a \lambda(u)S(u)e^{-\int_u^a \gamma dx} du = \int_0^a \lambda(u)S(u)e^{-\gamma(a-u)}du \end{aligned} \qquad (4.13)$$

Si trova cosi la seguente equazione integrale per la forza dell'infezione:

$$\begin{aligned} \lambda(a) &= \int_0^\infty q(a,x)C(a,x)I(x)dx = \int_0^\infty q(a,x)C(a,x)\left(\int_0^x \lambda(u)S(u)e^{-\gamma(x-u)}du\right)dx = \\ &= \int_0^\infty \int_0^x q(a,x)C(a,x)\lambda(u)e^{-\int_0^u \lambda(t)dt} e^{-\gamma(x-u)}dudx \end{aligned} \qquad (4.14)$$

L'analisi matematica della precedente equazione mostra come in effetti sia possibile dimostrare (Inaba, 1990) sotto opportune condizioni l'esistenza di almeno uno stato di equilibrio non banale con forza di infezione positiva, detto stato di *equilibrio endemico*. In particolare nel caso semplificato per classi di età discrete con forza di infezione in equilibrio:

$$\lambda_i = \sum_{j=1}^{m} q_{ij} C_{ij} \bar{I}_j \qquad (4.15)$$

(Dove abbiamo indicato con $\bar{I}_j$ la frazione infettiva di equilibrio nella classe di eta j-esima), la forma delle soluzioni del modello dipendono dall'autovalore dominante della matrice non-negativa (detta matrice dei casi di nuova generazione):

$$NG = \left[Dq_{ij} C_{ij} \right] \qquad (4.16)$$

Si indichi con R_0 tale autovalore dominante. Per forme non patologiche della matrice di nuova generazione si trova che se R_0 è minore od uguale ad uno allora l'equilibrio DFE è globalmente stabile, mentre se R_0 è maggiore di uno allora esiste uno stato di equilibrio endemico che è almeno localmente stabile (Inaba, 1990). Per forme "strane" della matrice di nuova generazione è stato recentemente dimostrato (Franceschetti e Pugliese, 2010) che possono comparire equilibri multipli. Tale autovalore dominante ha un' importante interpretazione epidemiologica, ossia rappresenta il *numero di riproduzione di base dell'infezione* (Anderson e May, 1991), definito come il numero di casi secondari che un soggetto infetto si aspetta di produrre nel corso del suo intero periodo di infettività in una popolazione interamente suscettibile.

5 Stima dei parametri di trasmissione

5.1 L'approccio tradizionale

La formulazione del problema matematico mostra degli evidenti problemi di eccessiva parametrizzazione. Se m è il numero di classi d'età considerate, i parametri complessivi da stimare per il modello sono un totale di $2m^2$ (m^2 parametri di contatto e m^2 coefficienti di trasmissione). L'approccio tradizionale "Who Aquires the Infection From Whom" (WAIFW) sviluppato, a partire da Anderson e May (1991) ed utilizzato fino ad anni molto recenti nelle applicazioni di sanità pubblica, aveva come radice l'idea di determinare i parametri incogniti mediante riduzione dello spazio parametrico fino alla sua riconduzione ad un sistema di equazioni esattamente identificato. Questo approccio, matematicamente molto conveniente. ma statisticamente grossolano, valido nel caso della disponibilità di dati-prevaccinali, era operazionalizzato come segue:

- Si assume che l'infezione sia nel suo stato di equilibrio endemico pre-vaccinale, consentendo così di utilizzare le equazioni di equilibrio del modello.
- Si stima la forza dell'infezione di equilibrio λ_i sulle varie classi di età mediante ordinarie tecniche di analisi di sopravvivenza per modelli con hazard costante a tratti, utilizzando la versione semplificata SR del modello. Si indichi con $\hat{\lambda}_i$ la stima ottenuta.
- nella ipotesi di considerare il medesimo numero di classi di età per il dato sierologico ed il dato di contatto : (a) si definivano i parametri compositi $C_{ij}^A = q_{ij} C_{ij}$, detti "contatti adeguati"; questa posizione riduce il numero di incognite da $2m^2$ ad m^2 ; (b) si introducono – mediante considerazioni di carattere teorico - restrizioni ad hoc sui parametri della matrice dei contatti adeguati al fine di ridurre il numero di incognite da m^2 ad m, rendendo quindi il sistema esattamente identificato.
- Una volta disponibile una stima della forza dell'infezione, il sistema di equilibrio viene risolto matematicamente, ottenendo il sistema di equazioni lineari nelle incognite C_{ij}^A :

$$\hat{\lambda}_i = \sum_{j=1}^{m} C_{ij}^A \bar{I}_j\left(\hat{\lambda}_i\right) \qquad (5.1)$$

che può in generale essere risolto per i contatti adeguati in maniera esatta.

L'ovvio limite dell'approccio WAIFW sta nella sua grossolanità statistica: si riduceva la dimensione dello spazio parametrico mediante ipotesi e senza utilizzare il dato. Inoltre per ottenere l'uguaglianza tra numero di classi di età del dato sierologico e del dato di contatto, non si faceva un uso appropriato del dato sierologico, sovente disponibile su scale di età molto (annuali).

5.2 Stima dei parametri in presenza di dati di contatto

La disponibilità del dato di contatto offre ovviamente una prospettiva completamente nuova rispetto all'approccio WAIFW. Naturalmente occorrerebbe verificare che la tipologia di contatto considerata sia appropriata per l'infezione in esame. Fortunatamente il dato Polymod offre una vasta tipologia di possibili contatti cosi che la minore o maggiore appropriatezza di differenti tipologie di contatti può essere direttamente sottoposta a verifica nel corso della procedura di adattamento. Assumendo di disporre della matrice di contatto "appropriata", indicata semplicemente con C_{ij} , le uniche incognite del problema rimangono i coefficienti di trasmissione. Assumendo inoltre un numero ridotto *s* di coefficienti di trasmissione incogniti $(q_1,..,q_s)$ (la ovvia motivazione è che ci aspettiamo che solo alcune delle classi di età su cui si verificano i contatti siano caratterizzate da contagiosità specifiche

e distinte tra loro) ed un numero *K>>s* di classi di età del dato sierologico otteniamo un problema di stima non lineare con *(K-s)* gradi di libertà, stimabile con una tecnica di verosimiglianza per dati sierologici. Nel seguito sviluppiamo l'idoneo modello statistico.

5.3 Il modello statistico : la likelihood sierologica

Sia n la numerosità campionaria. Lo stato immunologico del generico individuo ad una qualunque età è rappresentabile tramite una variabile casuale di Bernoulli:

$$Y_i = \begin{cases} 1 & se \quad immune \\ 0 & se \quad suscettibile \end{cases}, \qquad (5.2)$$

con i=1….n . Consideriamo la generica classe d'età k del dato sierologico (k=1…K, K>m), ove nk denota il numero di osservazioni ($\sum_{k=1}^{K} n_k = n$) e y_k il numero realizzato di "successi", ovvero il numero di soggetti immuni riscontrato nel campione. Indichiamo con π_k è la probabilità che il generico individuo nella classe d'età k-esima risulti immune. Ovviamente tale probabilità può essere calcolata mediante la prevalenza attesa dal modello matematico SIR valutato in equilibrio – visto come modello di rischio in base alla considerazioni delle sezioni precedenti. Tale probabilità dipende dai parametri incogniti $(q_1,..,q_s)$: $\pi_k = \pi_k(q_1,..,q_s)$. Assumendo che il dato sierologico costituisca un campione di osservazioni indipendenti e identicamente distribuite (IID) la funzione di verosimiglianza relativa alle osservazioni della classe di età k è una usuale verosimiglianza bernoulliana :

$$L_k(q_1,..,q_s) = \prod_{i=1}^{n_k} (\pi_k(q_1,..,q_s))^{y_k} (1-\pi_k(q_1,..,q_s))^{n_k - y_k}. \qquad (5.3)$$

La verosimiglianza complessiva per tutte le classi d'età è quindi espressa dalla seguente "catena" di verosimiglianze bernoulliane :

$$L = L(q_1,..,q_s) = \prod_{k=1}^{K} L_k(q_1,..,q_s) \qquad (5.4)$$

La log verosimiglianza infine è :

$$\log(L) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} [y_k \log \pi_k(q_1..q_s) + (n_k - y_k)\log(1-\pi_k(q_1..q_s))]. \qquad (5.5)$$

Al fine del calcolo di misure per valutare la bontà dell'adattamento del modello ai dati risulta utile definire la log-verosimiglianza del modello saturo. Essa è ottenuta assumendo un appropriato parametro bernoulliano per ogni singola classe di età, stimato mediante la corrispondete stima di massima verosimiglianza, ossia dalla frazione campionaria di immuni (che naturalmente può assumere anche i valori estremi 0 ed 1, cosa che si verifica frequentemente per le classi estremali):

$$\log(L) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \left[y_k \log \frac{y_k}{n_k} + (n_k - y_k)\log(1-\frac{y_k}{n_k}) \right]. \qquad (5.6)$$

5.4 Calcolo delle frequenze attese in funzione dei parametri incogniti

Nella espressione della verosimiglianza appaiono per ora i parametri di trasmissione da stimare in una forma soltanto implicita attraverso le probabilità $\pi_k = \pi_k(q_1,..,q_s)$. Queste probabilità possono però essere rese in forma esplicita in funzione della forza dell'infezione di equilibrio del modello. Questo richiede una appropriata risoluzione del modello matematico in equilibrio al fine di determinare le prevalenze attese di soggetti immuni, che interverranno nella funzione di verosimiglianza. Ripartiamo pertanto dal modello in equilibrio sotto ipotesi di assegnata matrice dei contatti C_{ij} che implica una forza dell'infezione (di equilibrio) costante a tratti sulle classi di età della matrice dei contatti:

$$\lambda(a)=\begin{cases}\lambda_i & a_{i-1}<a<a_i \\ 0 & altrove\end{cases} \qquad (5.7)$$

Assumiamo inoltre, per semplicità di esposizione, che esista un unico coefficiente di trasmissione distinto q. Vedremo poi rapidamente come l'analisi può essere estesa al caso di una pluralità di coefficienti di trasmissione distinti. Ne segue che la forza di infezione di equilibrio sulla classe di età i-esima è:

$$\lambda_i = q\sum_{j=1}^{m} C_{ij}\bar{I}_j \qquad (5.8)$$

Sotto tale assunzione possiamo trovare una soluzione formale alle ODE di equilibrio in funzione della forza dell'infezione assunta come data mediante risoluzione ricorrente, partendo dalla prima classe di età. Procediamo pertanto ricorsivamente risolvendo innanzitutto per la classe dei suscettibili, descritta dall'equazione $\dot{X}(a)=-\lambda(a)X(a)$. Ricordiamo che nella ipotesi di rischio di morte del tipo 1 è equivalente procedere lavorando sui numeri assoluti oppure sulle frazioni epidemiologiche per età: le seconde sono scalate, rispetto alle prime, per il numero B di nascite annue. Vale in generale:

$$X(a)=X_0 e^{-\int_0^a \lambda(a)da} \qquad (5.9)$$

Pertanto se consideriamo l'intervallo generico $[a_{i-1},a_i)$ abbiamo

$$X(a)=X_0 e^{-\int_0^a \lambda(a)da}=B_0 e^{-\left(\int_0^{a_{i-1}} \lambda(a)da\right)} e^{-\left(\int_{a_{i-1}}^{a_i} \lambda(a)da\right)}=X(a_{i-1})e^{-\lambda_i(a_i-a_{i-1})} \qquad a_{i-1}<a<a_i \qquad (5.10)$$

Seguono quindi le relazioni:

$$X(a)=\begin{cases} Be^{-\int_0^a \lambda_1 ds}=Be^{-\lambda_1 a} & 0<a<a_1 \\ Be^{-\lambda_1 a_1}e^{-\lambda_2(a-a_1)}=X(a_1)e^{-\lambda_2(a-a_1)} & a_1<a<a_2 \\ \vdots & \\ Be^{-\lambda_1 a_1}e^{-\lambda_2(a_2-a_1)}\ldots e^{-\lambda_{i-1}(a_{i-1}-a_{i-2})}e^{-\lambda_i(a-a_{i-1})}=X(a_{i-1})e^{-\lambda_i(a-a_{i-1})} & a_{i-1}<a<a_i \end{cases} \qquad (5.11)$$

In forma compatta possiamo riscrivere:

$$X(a)=X(a_{i-1})e^{-\lambda_i(a-a_{i-1})}=B\left(\prod_{j=1}^{i-1}e^{-\lambda_j(a_j-a_{j-1})}\right)e^{-\lambda_i(a-a_{i-1})} \qquad a_{i-1}<a<a_i \qquad (5.12)$$

I corrispondenti valori della frazione suscettibile alle varie età $S(a)=X(a)/n(a)$ sotto l'ipotesi del tipo 1 in cui: $n(a)=B$ sono immediatamente ottenute dividendo per B le precedenti espressioni. Si noti che tali espressioni in effetti rappresentano le frazioni suscettibili per qualunque forma della funzione di mortalità, almeno fintantoché la mortalità per età agisce in maniera omogenea sulle varie classi epidemiologiche, come implicito nelle nostre equazioni fondamentali. Pertanto:

$$x(a)=x(a_{i-1})e^{-\lambda_i(a-a_{i-1})}=\left(\prod_{j=1}^{i-1}e^{-\lambda_j(a_j-a_{j-1})}\right)e^{-\lambda_i(a-a_{i-1})} \qquad a_{i-1}<a<a_i \qquad (5.13)$$

La frazione suscettibile all'età è ovviamente la *funzione di sopravvivenza nello stato suscettibile* valutata all'età esatta *a*.
In particolare vale:

$$x(a_i) = x(a_{i-1})e^{-\lambda_i(a_i - a_{i-1})} = \prod_{j=1}^{i} e^{-\lambda_j(a_j - a_{j-1})} \qquad (5.14)$$

.
Quindi segue:

$$x(a_i) - x(a_{i-1}) = x(a_{i-1})e^{-\lambda_i(a_i - a_{i-1})} - x(a_{i-1}) = x(a_{i-1})\left(e^{-\lambda_i(a_i - a_{i-1})} - 1\right) \qquad (5.15)$$

ossia

$$x(a_{i-1}) - x(a_i) = x(a_{i-1})\left(1 - e^{-\lambda_i(a_i - a_{i-1})}\right) \qquad (5.16)$$

La precedente relazione è ben nota nell'analisi del rischio con rischi costanti, ed afferma che variazione assoluta nella frazione/probabilità di sopravvivenza su di un intervallo assegnato è data dal complemento ad 1 del rischio accumulato sull'intervallo stesso.
Procediamo adesso alla risoluzione per gli infettivi. Per le malattie di durata breve vale con ottima approssimazione la seguente relazione:

$$Y(a) \cong \frac{1}{\nu}\lambda(a)X(a) = D\lambda(a)X(a) \qquad (5.17)$$

Ove D è la durata dell'infezione, La precedente relazione afferma che la prevalenza assoluta di infettivi è pari al prodotto della incidenza di nuove infezioni (il termine $\lambda(a)X(a)$) moltiplicata per la durata della fase infettiva, ed è ovvia in un modello con durate fisse, ma vale in effetti in generale. Se inoltre la forza dell'infezione è costante a tratti allora il numero atteso di infettivi presenti nella classe i-esima è

$$Y_i = \int_{a_{i-1}}^{a_i} Y(a)da = \int_{a_{i-1}}^{a_i} D\lambda_i X(a)da = D\lambda_i X_i \qquad (5.18)$$

dove:

$$X_i = \int_{a_{i-1}}^{a_i} X(a)da. \qquad (5.19)$$

Rappresenta il numero atteso di suscettibili presenti nella classe i-esima. Integrando la precedente espressione otteniamo, sempre per recursione:

$$X_1 = \int_0^{a_1} X(a)da = \int_0^{a_1} B\exp(-\lambda_1 a)da = \frac{B}{\lambda_1}\int_0^{a_1}\lambda_1\exp(-\lambda_1 a)da = \frac{B}{\lambda_1}(1 - \exp(-\lambda_1 a_1)) \qquad (5.20)$$

e così via:

$$X_i = \int_{a_{i-1}}^{a_i} X(a)da = \int_{a_{i-1}}^{a_i} X(a_{i-1})e^{-\lambda_i(a - a_{i-1})}da = \frac{X(a_{i-1})}{\lambda_i}\int_0^{(a_i - a_{i-1})}\lambda_i e^{-\lambda_i u}du =$$

$$= \frac{X(a_{i-1})}{\lambda_i}\left(1 - e^{-\lambda_i(a_i - a_{i-1})}\right) \qquad (5.21)$$

$$= \frac{1}{\lambda_i}\left(X(a_{i-1}) - X(a_i)\right)$$

La formula precedente che assegna la frazione che è mediamente suscettibile in ogni classe di età, ci assegna quindi anche le prevalenze attese relative π_k per ogni classe di età, come la frazione complementare che non è suscettibile:

$$\pi_i = 1 - \frac{X_i}{n_i} = 1 - \frac{1}{\lambda_i}\left(X(a_{i-1}) - X(a_i)\right) \qquad (5.22)$$

Infatti, individui infettivi verrebbero caratterizzati da elevati titoli anticorpali, e quindi correttamente classificati nella classe complementare degli immuni.

Dalla formula (5.21) e dalla (5.18) segue subito:

$$Y_i = D\lambda_i X_i = D\left(X(a_{i-1}) - X(a_i)\right) \qquad (5.23)$$

Ripartendo dalla definizione generale della forza di infezione $\lambda_i = \sum_{j=1}^{n} qC_{ij}\frac{Y_j}{n_j}$ otteniamo:

$$\lambda_i = \sum_{j=1}^{n} qC_{ij}\frac{Y_j}{n_j} = \sum_{j=1}^{n} qC_{ij}\frac{D\left(X(a_{j-1}) - X(a_j)\right)}{n_j} \qquad (5.24)$$

Ricordando le espressioni per le $X\left(a_j\right)$**:**

$$X(a_j) = B\prod_{h=1}^{j} e^{-\lambda_j(a_h - a_{h-1})} \qquad (5.25)$$

Segue quindi:

$$\lambda_i = \sum_{j=1}^{n} qC_{ij}\frac{DB\left(\prod_{h=1}^{j-1} e^{-\lambda_j(a_h - a_{h-1})} - \prod_{h=1}^{j} e^{-\lambda_j(a_h - a_{h-1})}\right)}{Bh_j} \quad i = 1,..,n \qquad (5.26)$$

Ossia

$$\lambda_i = Dq\sum_{j=1}^{n}\frac{C_{ij}}{h_j}\left(\prod_{h=1}^{j-1} e^{-\lambda_j(a_h - a_{h-1})} - \prod_{h=1}^{j} e^{-\lambda_j(a_h - a_{h-1})}\right) \quad i = 1,..,n \qquad (5.27)$$

La precedente espressione definisce, per *q* fissato, un sistema esplicito di n equazioni non lineari per la determinazione della forza dell'infezione in equilibrio. Questo sistema svolge un ruolo fondamentale nella procedura di ottimizzazione della verosimiglianza in quanto dobbiamo assicurare che i valori ottimi dei coefficienti q corrispondano in effetti ad una forza dell'infezione in equilibrio, e quindi soddisfacente i vincoli (5.26).

Oss. L'estensione della procedura precedente al caso della stima di un vettore di parametri è immediata.

5.5 Stima di massima verosimiglianza dei parametri di trasmissione

La formulazione del problema di stima è a questo punto completa . Dato un problema generico 1-parametrico con singolo coefficiente di trasmissione q, cerchiamo il valore ottimo q^{ML} che rende massima la funzione di log-versomiglianza,

$$\log(L(q)) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left[y_k \log \pi_k(q) + (n_k - y_k)\log(1-\pi_k(q))\right]. \qquad (5.28)$$

subordinatamente alla matrice dei contatti selezionata C_{ij}, e al fatto che la corrispondente forza dell'infezione soddisfaccia le condizioni di equilibrio;

$$\lambda_i = Dq\sum_{j=1}^{n}\frac{C_{ij}}{h_j}\left(\prod_{h=1}^{j-1}e^{-\lambda_j(a_h-a_{h-1})} - \prod_{h=1}^{j}e^{-\lambda_j(a_h-a_{h-1})}\right) \qquad i = 1,..,n \qquad (5.29)$$

Il problema di ottimizzazione della verosimiglianza deve essere risolto numericamente, non avendo soluzione analitica esplicita nemmeno nel caso 1-parametrico. In conseguenza delle condizioni di equilibrio (5.29) il problema risulta essere quindi un problema di ottimo vincolato. E' stata adottata la seguente procedura. Il problema di ottimizzazione della verosimiglianza è stato risolto mediante la routine numerica fminsearch di matlab, basata sull'algoritmo del simplesso nonlineare (descritto sinteticamente nella prossima sotto-sezione). Ogni volta che la routine fminsearch "campiona" un valore del parametro – quindi un candidato a rappresentare la stima di MV cercata - occorre chiedere che il vincolo (5.29) sia soddisfatto. Per fare questo sfruttiamo il fatto che la (5.29) stessa definisce un formula ricorrente di un problema di punto fisso stabile (se vale la condizione di esistenza di unico stato endemico). Pertanto è sufficiente assegnare un valore iniziale alla forza di infezione in ogni classe d'età: dato il valore q prescelto la (5.29) calcola allora iterativamente un nuovo vettore di valori della forza dell'infezione. Se tale vettore soddisfa la condizione di arresto (ossia la differenza tra i due membri dell'equazione risulta inferiore ad un valore prestabilito molto prossimo a zero, ad es:10^{-15}), allora tale vettore è accettato come forza dell'infezione di equilibrio corrispondente al valore q prescelto, condizionalmente alla matrice dei contatti. Altrimenti, se la condizione di arresto non è soddisfatta, l'algoritmo viene iterato, fino a soddisfacimento della condizione di arresto. Numericamente è stato incluso un opportuno ciclo *while* nella routine di ottimizzazione della verosimiglianza, al fine di tenere persistentemente sotto controllo le equazioni di equilibrio.

5.6 Algoritmo del simplesso

La routine numerica fminsearch di matlab utilizzata per ottimizzare la funzione di verosimiglianza è basata sull'algoritmo del simplesso nonlineare **di Nelder-Mead (Lagarias et al., 1990)**.
Si tratta di un metodo di ricerca diretta che non fa uso di gradienti numerici o analitici. Se n è la lunghezza del vettore parametrico incognito, un simplesso in uno spazio n-dimensionale è l'iper-triangolo caratterizzato da n+1 vettori distinti che sono i suoi vertici. Geometricamente, in uno spazio bidimensionale, un simplesso è un triangolo, nello spazio tridimensionale, è una piramide. Ad ogni passo della ricerca, un nuovo punto all'interno o in prossimità del simplesso corrente viene generato. Il valore della funzione nel nuovo punto è confrontato con i valori della funzione nei vertici del simplesso e, di solito, uno dei vertici è sostituito dal nuovo punto, dando un nuovo simplesso. Questo passo è ripetuto fino a quando il diametro del simplesso è minore della tolleranza specificata. Un limite di tale metodo è che il minimo trovato potrebbe però essere soltanto locale.

5.7 Sommario: fasi dell'inferenza in modelli di trasmissione

La Fig. 5.1 riporta le fasi tipiche dello sviluppo delle procedure inferenziali per la stima dei coefficienti di trasmissione per il modello epidemiologico di base.

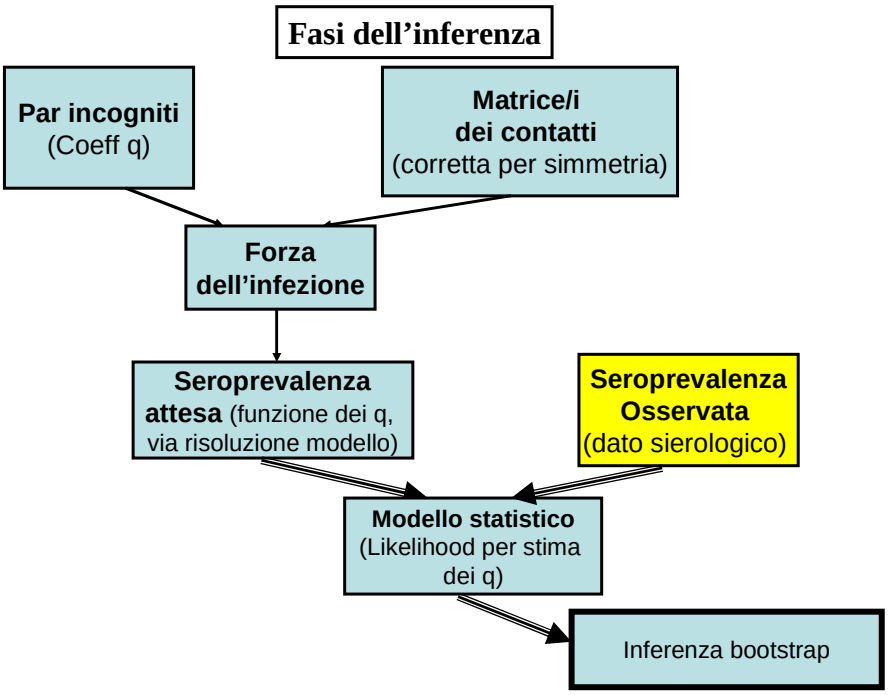


Fig. 5.1:Diagramma delle fasi dell'inferenza sui parametri di trasmissione

In breve sintesi: una volta scelta la matrice dei contatti, la forza di infezione d'equilibrio del modello è funzione del vettore di parametri q specifici per età. La corrispodente seroprevalenza attesa (dalla soluzione del modello matematico in equilibrio), è quindi espressa in funzione dei coefficienti di trasmissione. Dai dati sierologici,viene calcolata la seroprevalenza osservata. La stima del valore dei coefficienti di trasmissione si ottiene massimizzando la verosimiglianza statistica dei dati di sierologia che compara le seroprevalenze attese con le seroprevalenze osservate.
Sulle stime dei parametri è possibile fare inferenza o attraverso tecniche di ricampionamento (bootstrap) o tramite il concetto di verosimiglianza profilo.

6 Contatti sociali

6.1 Misurazione diretta dei contatti sociali

L'approccio diretto alla osservazione e misurazione dei contatti sociali mediante survey campionarie del comportamento sociale, è relativamente recente. Il primo vero studio diretto è lo studio, molto vasto, ideato e condotto in Olanda dal team di Van Druten (Wallinga et al., 2006) a partire dagli anni 80 del secolo passato. Sorprendentemente i dati prodotti da questo studio sono rimasti completamente inutilizzati, almeno dal punto di vista epidemiologico, fino al lavoro di Wallinga et al. (2006) che sviluppa l'approccio di base alla stima dei coefficienti di trasmissione mediante combinazione di dati di contatto e dati di sierologia. La limitazione del dataset adottato da Wallinga è dovuta principalmente alla semplicità del dato: si riportano soltanto i numeri delle conversazioni che i partecipanti hanno avuto con persone di differenti età durante una settimana tipica. Essendo lo studio olandese passato completamente sotto silenzio, i primi studi noti sull'approccio diretto alla misurazione dei contatti sono quelli sviluppati da Edmunds a partire dalla fine degli anni 1990 (Edmunds et al., 1997) mediante diari di contatto in popolazioni universitarie, cui hanno fatto seguito i contributi di Beutels et al. (2003), Wallinga et al. (2006), Edmunds et al. (2006). Questi lavori costituiscono il punto di partenza del progetto FP6 Polymod, coordinato da J. Edmunds, che coinvolge in un unico team di ricerca tutti gli sviluppatori dell'approccio.

6.2 Metodologia dell'indagine Polymod sui contatti sociali

Il progetto Polymod è un'indagine trasversale condotta contemporaneamente in otto Paesi Europei (Belgio (BE), Germania (DE), Finlandia (FI), Gran Bretagna (GB), Italia (IT), il Lussemburgo (LU), Paesi Bassi (NL), e Polonia (PL).), dopo una lunga preparazione mirante ad individuare un insieme appropriato di definizioni di marcatore di contatto ed a creare un diario di contatti "sufficientemente" standardizzato ed armonizzato, da rendere cosi possibili comparazioni internazionali. La dizione "sufficientemente" è inevitabile per le difficoltà incontrata nel rendere appropriatamente nelle diverse lingue concetti non immediatamente ovvi come "avere avuto un contatto fisico con qualcuno".

Il recrutamento e la raccolta dei dati sono stati organizzati a livello nazionale secondo i pesi campionari concordati e l' utilizzo comune di diari per le registrazioni dei contatti. Il campione è rappresentativo dell'intera popolazione, in termini di diffusione geografica, età e sesso (ma con un sovra campionamento delle classi di età giovanili).

Le indagini sono state materialmente svolte tra Maggio2005 e Settembre 2006. I partecipanti al sondaggio sono stati reclutati telefonicamente usando la tecnica RDD ('*Random Digit Dialing*' o attraverso i registri della popolazione. Bambini ed adolescenti sono stati deliberatamente sovracampionati, dato l'importante ruolo che rivestono nella diffusione di agenti infettivi. Inoltre, per ogni nucleo famigliare un solo componente ha partecipato all'indagine. Le agende in carta sono state inviate tramite mail o consegnate direttamente ai partecipanti. Si definisce contatto una conversazione con uno o più interlocutori (contatto non fisico) o un contatto fisico (skin to skin) come un bacio o una carezza. Sui diari vengono registrate *informazioni socio-demografiche di base* sui partecipanti, comprensive di occupazione, livello di istruzione, la composizione del nucleo famigliare, l'età e il sesso. Ad ogni partecipanti viene indicato un giorno a caso della settimana per registrare i contatti avuti con altre persone tra le 05:00 a.m. e le 05:00 a.m. della mattina seguente, registrando ciascun individuo una sola volta nel diario. Ai partecipanti è stato anche chiesto di fornire informazioni circa l'età e il sesso di ogni persona contattata. Quando l'età del contatto non è nota precisamente, i partecipanti stimano una banda d'età (nell'analisi dei dati si utilizza il punto centrale di tale banda come stima puntuale dell'età del contatto). Per ogni contatto, i partecipanti hanno registrato la prossimità ,il luogo (casa, scuola, lavoro, tempo libero, trasporto, altro), la durata (meno di 5 min, 5–15 min,15 min - 1 h, 1–4 h, più di 4 h) e la frequenza usuale del contatto

(giornaliera, settimanale, circa 1 o 2 volte al mese, meno di una volta al mese, per la prima volta). I diari sono tradotti e disponibili in varie lingue. Per evitare degli errori nella compilazioni sono state utilizzate 3 forme distinte di diari differenziando tra bambini(6-10 anni),adolescenti(11-18) e adulti. In diari dei neonati e dei bambini più piccoli sono stati compilati dai genitori .

6.3 Risultati di base dell'indagine

Ciò ha reso possibile investigare con dettaglio le caratteristiche dei contatti ,utilizzando l'età (in primis) e molte altre covariate per tale scopo. Attraverso questo ricco dataset si sono potute stimare le matrici dei numeri medi di contatto per giorno, suddivisi per età del partecipante e del contatto, per varie tipologie di contatti direttamente. In totale a livello europeo sono stati coinvolti nello studio n=7,290 partecipanti, i quali hanno registrato le caratteristiche di 97,904 contatti con differenti individui durante un giorno della settimana, includendo età, sesso, luogo, durata, frequenza ed occorrenza del contatto fisico. Le distribuzioni e le caratteristiche dei contatti per età sono simili per i differenti Paesi Europei. I contatti sociali risultano altamente assortativi con l'età: in particolare i bambini e gli adolescenti tendono ad avere contatti con persone della stessa età.
Osservando le superfici dei contatti(Mossong, 2007) appare evidente che il maggior numero di contatti avviene tra individui appartenenti alla stessa classe d'età(la diagonale principale è molto frequentata). Tale tendenza è più marcata tra 5 e 24 anni e meno evidente tra 55 e 69 anni. Inoltre le due diagonali secondarie che iniziano sia per i partecipanti sia per i contatti tra 30 e 35 anni evidenziano i contatti prevalentemente in famiglia: sono i contatti tra genitori e figli.

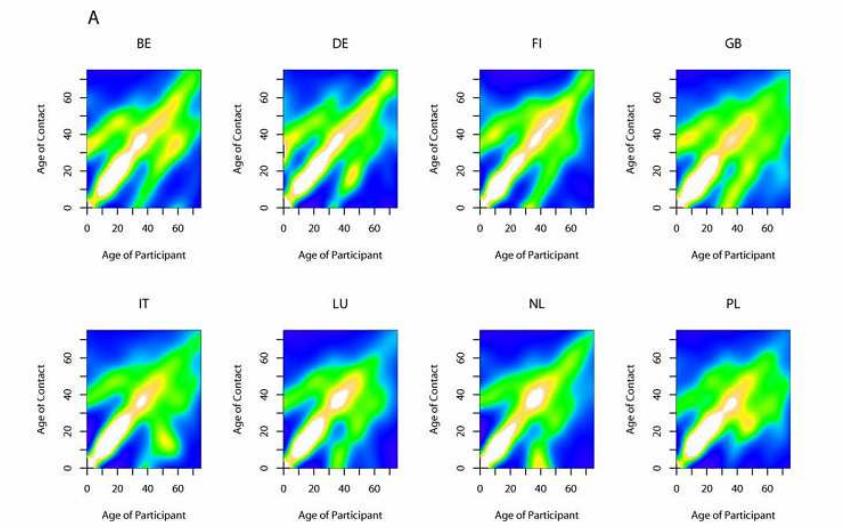


Fig. 6.1:Matrice di contatto 'Smooth' per gli 8 paesi dell'indagine Polymod

I contatti che durano almeno un' ora e che avvengono quotidianamente quasi sempre sono di tipo fisico, mentre i contatti di breve durata e sporadici sono per lo più non fisici. I contatti in famiglia, a scuola e nel tempo libero sono spesso fisici. Durante un viaggio od utilizzando i trasporti pubblici e sul posto di lavoro i contatti sono solitamente non fisici. I modelli matematici hanno evidenziato che la fascia d'età compresa tra 5 e 19 anni è mediamente quella con la maggiore incidenza in caso di una pandemia trasmessa tramite i contatti sociali, quando la popolazione è completamente suscettibile (Mossong et al. ,2008).

6.4 Costruzione delle matrici di contatto sociale

In questa sezione presentiamo la procedura con cui, a partire dai dati grezzi dell'indagine Polymod, vengono costruite le matrici di contatto sociale per età, che verranno utilizzate come base del modello matematico SIR, e del corrispondente modello statistico per la stima dei coefficienti di trasmissione per la varicella in Italia. Notiamo preliminarmente che analizzando i dati per un singolo paese non è necessario considerare il peso statistico assegnato, nell'indagine complessiva, ad ognuno dei paesi partecipanti.
Un punto preliminare di una certa importanza relativo alla utilizzazione delle matrici di contatto era se fosse preferibile servirsi delle matrici grezze, come emergono direttamente dai dati, oppure se procedere a procedure di lisciamento ("smoothing") mediante idonei modelli, al fine di conferire alle matrici utilizzate delle forme più regolari. Abbiamo deciso alla fine di servirci delle matrici "grezze", al fine di non intaccare in minima misura l'informazione empirica. Pertanto non faremo alcuna assunzione sulle distribuzioni congiunte dei contatti al variare dell'età. Ovviamente il passo successivo non può che essere la ricerca di criteri di smoothing in grado di tenere conto appropriatamente delle struttura interne delle matrici di contatto.
La costruzione delle matrici grezze prevede come primo passo la creazione di idonee classi di età discrete, del tipo $[a_{i.1}, a_i)$, e successivamente la assegnazione di ogni singolo contatto registrato (per tutti i partecipanti, per ogni loro contatto) ad una classe di età congiunta, in base alle classi d'età del partecipante e dei suoi contatti[1]. In tabella 1 vengono presentate le più importanti variabili investigate nell'indagine Polymod.

Tabella 6.1Variabili per il contatto (principali) dell'indagine Polymod

Nome delle Variabili	Etichetta
Country *local_id* *global_id*	*Variabili identificative del paese e del diario*
cnt_count	*Numero di contatti nel diario*
cnt_age_mean	*Età media*
cnt_ age_l	*Estremo inferiore dell'età del contatto*
cnt_age_r	*Estremo superiore dell'età del contatto*
cnt_sex	*Genere del contatto*
cnt_home	*Contatti a casa*
cnt_work	*Contatti a lavoro*
cnt_school	*Contatti a scuola*
cnt_transport	*Contatti durante gli spostamenti*
cnt_leisure	*Contatti durante le attività del tempo libero*
cnt_otherplace	*Contatti durante altre attività*
cnt_frequency	*Frequenza del contatto*
cnt_touch	*Contatti fisici*
cnt_duration	*Durata totale del contatto durante l'intero giorno*

[1] Il dato viene censurato se l'età del contatto o del partecipante è maggiore uguale ad 80 anni, dato che l'età massima registrata nei dati di sierologia è 79 anni.

In questo modo si ottiene la distribuzione assoluta congiunta dei contatti totali giornalieri (ovvero contatti di tutte le tipologie), in base alle età dei partecipanti e dei loro contatti. Naturalmente la procedura può essere ripetuta considerando anziché tutti i contatti registrati, solo una o alcune selezionate tipologie di contatti, come possono essere per esempio i contatti di tipo "fisico". Il vantaggio del dato Polymod è che fornisce informazioni su una vasta serie di tipologie di contatti, dei quali è di grande interesse andare a verificare, mediante un criterio di adattamento statistico, quali siano quelli che meglio spiegano i dati della trasmissione della varicella. Infatti, a tutt'oggi, non è noto per alcuna infezione a trasmissione per contatto diretto, quali siano le forme di contatto a rischio più efficaci tra due individui al fine di determinare la trasmissione dell'infezione, Da questo punto di vista l'utilizzo del dato Polymod può fornire un contributo molto importante ad antichi problemi irrisolti dell'epidemiologia delle infezioni.
Indichiamo pertanto con M_{ij} il numero di contatti totali giornalieri (ma il discorso può essere ripetuto senza modifiche per una qualunque tipologia specifica di contatti) intercorsi nel campione tra partecipanti della classe di età j-esima con soggetti della classe di età i-esima, e con x_j il numero di partecipanti della classe j. Dalla matrice dei contatti totali otteniamo quindi immediatamente la matrice dei contatti medi giornalieri per rispondente nel campione, che ha per elementi le quantità:

$$m_{ij} = M_{ij} / x_j \qquad (6.1)$$

Dalla matrice $[m_{ij}]$ otteniamo le stime appropriate dei numeri medi di contatti imponendo il vincolo di simmetria dei contatti. Tutte le definizioni di "contatto a rischio" utilizzate nell'indagine Polymod sono infatti simmetriche. Si prenda per esempio il marcatore "two-ways conversation": se per esempio il rispondente "signor A" ha dichiarato di avere avuto una conversazione con il suo contatto "B", allora necessariamente B ha avuto una conversazione con A. Si noti che questo non è rilevabile dal dato empirico nemmeno nel caso – per altro straordinariamente improbabile - in cui anche "B" fosse un rispondente selezionato nel campione. Per contatti di tipo simmetrico valgono, in una popolazione stratificata per età e costituente una rete "chiusa" di contatti (in cui cioè tutti i contatti sono all'interno della popolazione medesima) le relazioni:

$$K_{ij} = C_{ij} w_j = C_{ji} w_i = K_{ji} \qquad (6.2)$$

Dove w_j è la numerosità assoluta della popolazione nella classe di età j-esima, C_{ij} l'esatto numero medio di contatti age-specific della popolazione, e $K_{ij} = K_{ji}$ il corrispondente numero totale di contatti tra soggetti di età i e j nella popolazione. Naturalmente le stime campionarie m_{ij} non possono soddisfare il vincolo di simmetria, in quanto il campione osservato di rispondenti per definizione non costituisce una rete chiusa di contatti. Tuttavia possiamo correggere le stime mediante il principio di simmetria al fine di renderle consistenti con l'ampiezza demografica osservata della popolazione di riferimento. A tale scopo calcoliamo i contatti totali attesi di popolazione mediante le loro stime campionarie. Troviamo così le stime $\hat{K}_{ij} = m_{ij} w_j$ e $\hat{K}_{ji} = m_{ji} w_i$ che in generale non soddisferanno il vincolo di simmetria, ovvero: $\hat{K}_{ij} \neq \hat{K}_{ji}$. Sembra a questo punto ragionevole forzare la simmetrizzazione mediante ricorso ad opportune medie tra tali quantità. Per esempio, utilizzando la media aritmetica semplice dei K otteniamo la stima (simmetrica):

$$K_{ij,s} = \frac{\hat{K}_{ij} + \hat{K}_{ji}}{2} = K_{ji,s} \qquad (6.3)$$

A questo punto possiamo calcolare le stime corrette per la simmetria dei numeri medi di contatti come:

$$m_{ij,s} = \frac{K_{ij,s}}{w_j} = \frac{\hat{K}_{ij} + \hat{K}_{ji}}{2w_j} \qquad (6.4)$$

Si noti che mentre la matrice dei contatti totali attesi corretta per simmetria $K_s = [K_{ij,s}]$ è simmetrica, ovviamente la matrice dei contatti medi corretti $m_{ij,s}$ non è in generale mai simmetrica perché riflette gli sbilanci demografici tra i due gruppi di età i e j: se il gruppo di età j è più numeroso del gruppo i allora necessariamente avrà un numero medio di contatti inferiore.

6.5 Italia: matrici dei contatti medi per diverse tipologie di contatto

Stabilite le classi d'età dei contatti e dei partecipanti ,è possibile creare diverse matrici dei numeri medi dei contatti sociali . Le classi d'età dei partecipanti (j) e dei contatti(i) scelte per l'analisi dei dati Italiani sono:**0-1, 2-3, 4-10, 11-16, 17-21, 22 -24, 25-34, 35-44, 45-64, e 65-79**. Tali classi di età sono state scelte al fine di poterne comparare i risultati con altri lavori in corso sullo stesso argomento (Melegaro et al., 2011, under review), e riflettono innanzi tutto l'importanza delle classi giovanili distinte per grado scolare e anche per differenti livelli di autonomia all'interno di un dato grado scolare.
Le correzioni per simmetria sono state effettuate utilizzando il dato Istat della popolazione per classi di età (Istat 2008).
Le tipologie di contatti che il dataset Polymod fornisce direttamente e indirettamente sono molte. Oltre ai contatti totali ed alle tipologie direttamente osservate di contatti (esempio: contatti stratificati *per prossimità*: fisici/non fisici, per *durata del contatto*: minore di 15 minuti, tra 15 e 60 minuti, maggiore di 60 minuti, per *luogo del contatto*: a casa, a lavoro, a scuola ,in altri luoghi, *per frequenza del contatto*: giornaliera,con frequenza settimanale, occasionali) se ne possono costruire molte altre potenzialmente interessanti intersecando tali caratteristiche (per esempio: contatti fisici di lunga durata, se si ritiene che la trasmissione debba richiedere soprattutto contatti intensi nel senso di ravvicinati, e prolungati nel tempo).
Di seguito vengono mostrate alcune matrici del numero medio di contatti giornalieri,quando si considerano tipologie specifiche di contatto (esempio:contatti fisici,contatti con durata minore di 15 minuti,contatti in famiglia,contatti occasionali).
La matrice dei contatti medi complessivi (Tabella 6.2) rivela una caratteristica fondamentale dei contatti sociali; la *assortatività*. Per assortatività (Hethcote, 1995) si intende la caratteristica ,osservata in molti fenomeni differenti,per cui gli individui tendono ad interagire socialmente con soggetti di caratteristiche simili alle loro. Nei nostri esempi, dove l'unico fattore fondamentale di stratificazione è l'età, l'interazione sociale tende ad indirizzarsi soprattutto verso individui della medesima età, questo si rileva molto chiaramente ispezionando la matrice della tabella 6.1: gli elementi della diagonale principale della matrice sono sistematicamente più elevati (con differenze maggiori fino a due/tre ordini di grandezza) degli elementi non diagonali della colonna corrispondente.
Per esempio soggetti della quarta classe d'età (11-16) risultano avere più di 16 contatti giornalieri(in media) con soggetti di pari età, mentre hanno meno di 3 contatti medi al giorno con soggetti di età immediatamente più grande, e praticamente non ne hanno(quasi 500 volte di meno!) del tutto con bambini della prima classe d'età. D'altra parte hanno quasi 3 contatti in media con soggetti della classe 35-44, rappresentati verosimilmente dai contatti con genitori e con qualche

altro adulto, verosimilmente in ambiente scolastico. Come vediamo quindi la matrice dei contatti ci informa molto accuratamente sulla distribuzione che la struttura *socio-istituzionale* (famiglia, scuola, etc) impone alle attività sociali di una comunità. La presenza di assortatività può essere misurata per mezzo di comuni misure statistiche quali il coefficiente di correlazione lineare di Bravais-Person oppure la dissomiglianza calcolate sulla distribuzione congiunta dei contatti totali (Farrington et al., 2009).

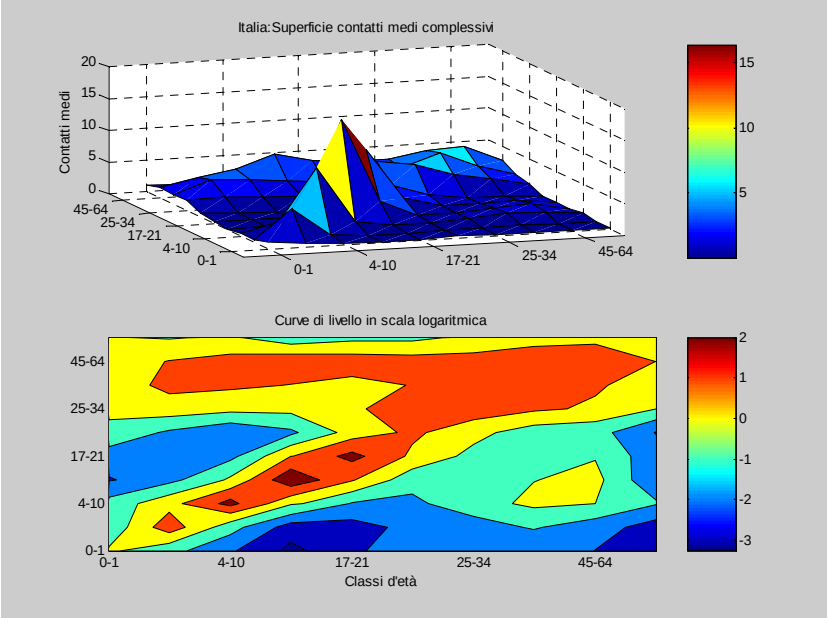


Fig. 6.2. .Superficie contatti medi complessivi e curve di livello in scalo logaritmica

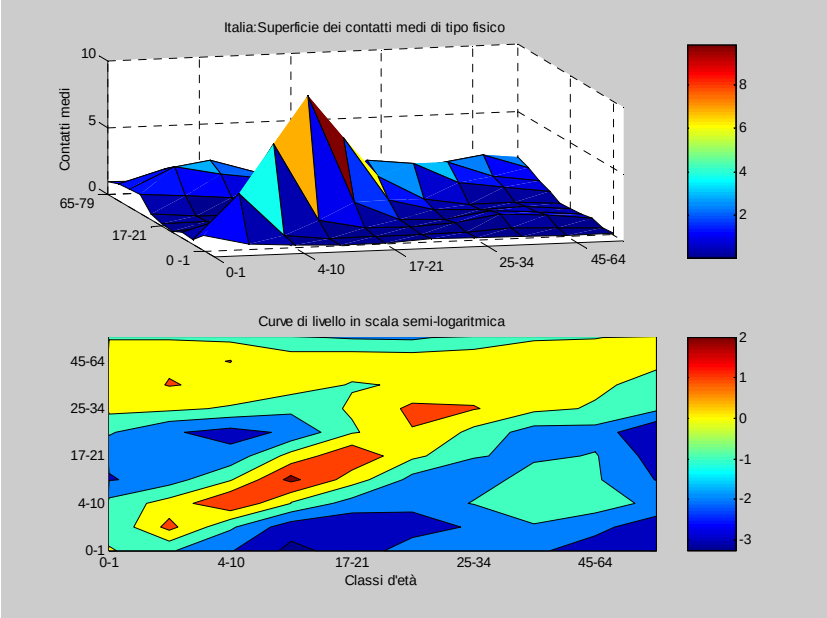


Fig. 6.3. Italia.Superficie contatti medi fisici e curve di livello in scalo logaritmica

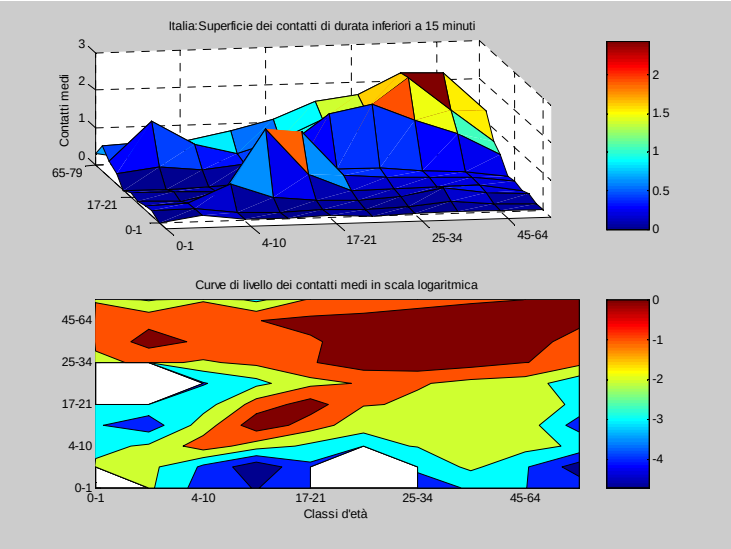


Fig. 6.4.Italia. Superficie contatti medi con durata inferiori a 15 minuti e curve di livello in scalo logaritmica

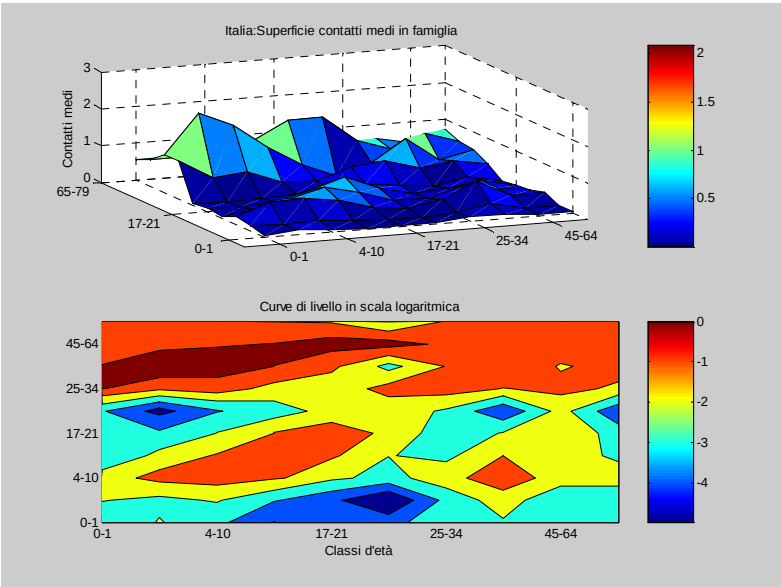


Fig. 6.5.Italia. Superficie contatti medi in famiglia e curve di livello in scalo logaritmica

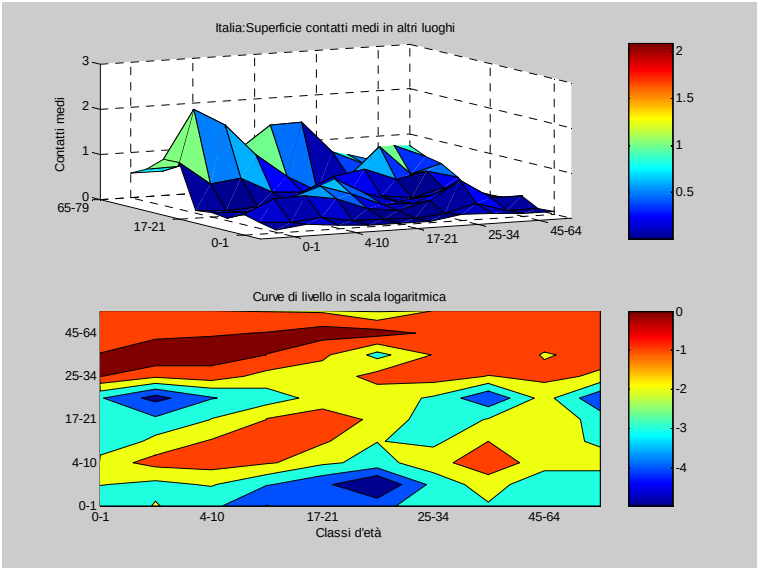


Fig. 6.6.Italia. Superficie contatti medi in altri luoghi e curve di livello in scalo logaritmica

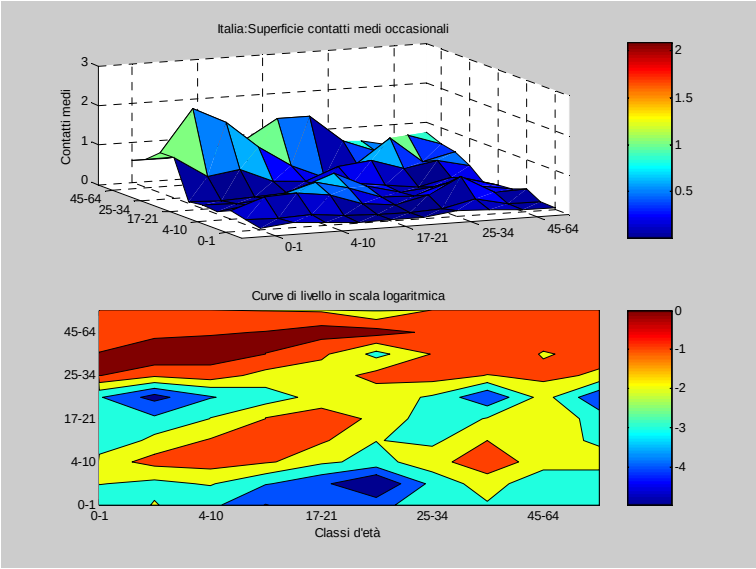


Fig. 6.7. Italia. Superficie contatti medi occasionali e curve di livello in scalo logaritmica

Le tabelle esposte in successione sono le matrici dei contatti medi italiani calcolate per diverse tipologie del contatto : nella cella (i,j) è espresso il numero medio di contatti che i partecipanti della classe d'età j-esima hanno con individui della classe d'età i-esima (con i,j=1…10) per la tipologia di contatto considerata.

i,j	1	2	3	4	5	6	7	8	9	10
Classe d'età	0-1	2-3	4-10	11-16	17-21	22-24	25-34	35-44	45-64	65-79

Tabella 6.2. ITALIA. Indagine Polymod: Matrice dei contatti medi complessivi

1.1579	**0.5112**	**0.15495**	**0.038719**	**0.10252**	**0.32417**	**0.21156**	**0.18854**	**0.13487**	**0.10465**
0.50266	**4.9487**	**0.55098**	**0.08481**	**0.08933**	**0.18269**	**0.23609**	**0.40429**	**0.20692**	**0.087551**
0.53085	**1.9198**	**10.0431**	**1.2732**	**0.34686**	**0.29908**	**0.64854**	**1.2699**	**1.0082**	**0.40333**
0.12127	**0.27017**	**1.164**	**16.3604**	**3.0494**	**0.52509**	**0.54646**	**0.97447**	**1.3748**	**0.15274**
0.27521	**0.24389**	**0.27179**	**2.6135**	**10.2034**	**2.2208**	**0.74284**	**0.66106**	**0.92757**	**0.18993**
0.58283	**0.33407**	**0.15696**	**0.3014**	**1.4873**	**3.3438**	**1.3783**	**0.59267**	**0.5512**	**0.13049**
1.6447	**1.8667**	**1.4717**	**1.3563**	**2.1512**	**5.9597**	**5.1591**	**3.3148**	**2.3061**	**1.0251**
1.6093	**3.5097**	**3.1639**	**2.6555**	**2.1019**	**2.8137**	**3.6395**	**5.6957**	**3.2018**	**1.5208**
1.8051	**2.8168**	**3.9387**	**5.8746**	**4.6247**	**4.1034**	**3.9703**	**5.0207**	**5.4151**	**2.7114**
1.0805	**0.91933**	**1.2155**	**0.50346**	**0.73046**	**0.74934**	**1.3614**	**1.8395**	**2.0915**	**2.5**

Tabella 6.3. ITALIA. Indagine Polymod Matrice dei contatti fisici

1.1579	**0.4064**	**0.14633**	**0.03871**	**0.07311**	**0.23141**	**0.1777**	**0.1533**	**0.0947**	**0.09445**
0.39961	**3.7436**	**0.4305**	**0.075801**	**0.066768**	**0.081931**	**0.15415**	**0.3313**	**0.1678**	**0.08511**
0.50132	**1.5**	**6.819**	**0.98612**	**0.23758**	**0.1865**	**0.39446**	**0.8707**	**0.70608**	**0.26288**
0.12127	**0.24147**	**0.90159**	**9.8468**	**1.7124**	**0.27438**	**0.19168**	**0.6407**	**0.41295**	**0.13088**
0.19626	**0.18229**	**0.18617**	**1.4676**	**6.0847**	**1.2928**	**0.39803**	**0.2902**	**0.3984**	**0.08519**
0.41605	**0.14982**	**0.097876**	**0.1575**	**0.86582**	**2.5625**	**0.76997**	**0.2304**	**0.2446**	**0.05165**
1.3815	**1.2188**	**0.89512**	**0.47574**	**1.1527**	**3.3294**	**2.9545**	**1.3025**	**0.9016**	**0.44981**
1.309	**2.8765**	**2.1696**	**1.746**	**0.92273**	**1.0942**	**1.4301**	**2.3652**	**1.2028**	**0.72252**
1.2683	**2.2843**	**2.7585**	**1.7646**	**1.9864**	**1.8213**	**1.5524**	**1.8861**	**2.3915**	**1.2141**
0.9752	**0.89369**	**0.79222**	**0.43139**	**0.32766**	**0.29663**	**0.59737**	**0.87393**	**0.93649**	**1.5652**

Tabella 6.4.ITALIA. Indagine Polymod Matrice dei contatti non fisici

0	**0.10481**	**0.00862**	**0**	**0.0294**	**0.0781**	**0.0338**	**0.0320**	**0.0401**	**0.00764**
0.10305	**0.20513**	**0.12048**	**0.00900**	**0.0225**	**0.0937**	**0.0632**	**0.0554**	**0.0391**	**0.00244**
0.02953	**0.41979**	**2.8448**	**0.25981**	**0.1092**	**0.1125**	**0.2446**	**0.3652**	**0.2715**	**0.11957**
0	**0.0287**	**0.23754**	**5.973**	**1.259**	**0.2428**	**0.3475**	**0.3161**	**0.8155**	**0.021865**
0.078947	**0.061599**	**0.085627**	**1.079**	**4.0508**	**0.88411**	**0.34482**	**0.3602**	**0.4811**	**0.0785**
0.14046	**0.17143**	**0.059083**	**0.1394**	**0.59213**	**0.78125**	**0.59179**	**0.34587**	**0.2866**	**0.0788**
0.26316	**0.50032**	**0.55505**	**0.86256**	**0.99858**	**2.5589**	**2.125**	**1.9394**	**1.303**	**0.5275**
0.27395	**0.48144**	**0.90999**	**0.86162**	**1.1453**	**1.642**	**2.1293**	**3.1304**	**1.8461**	**0.7005**
0.53682	**0.53243**	**1.061**	**3.4849**	**2.4012**	**2.1337**	**2.2433**	**2.8948**	**2.7594**	**1.299**
0.078947	**0.025641**	**0.36033**	**0.072072**	**0.30224**	**0.45271**	**0.70056**	**0.84732**	**1.002**	**0.69565**

Tabella 6.5. ITALIA. Indagine Polymod Matrice dei contatti medi per durata inferiore a 15 minuti

0	**0.1583**	**0.03166**	**0.01741**	**0.04768**	**0**	**0.03723**	**0.05312**	**0.05072**	**0.03126**
0.15568	**0.17949**	**0.043629**	**0.009009**	**0.039511**	**0**	**0.055987**	**0.17162**	**0.055614**	**0.01098**
0.1084	**0.1520**	**0.62069**	**0.23456**	**0.10928**	**0.0879**	**0.14606**	**0.3393**	**0.19572**	**0.1129**
0.0545	**0.0287**	**0.21445**	**1.8919**	**0.64199**	**0.1802**	**0.18189**	**0.1983**	**0.22875**	**0.0314**
0.1280	**0.1078**	**0.08562**	**0.55021**	**1.6102**	**0.4033**	**0.31882**	**0.2838**	**0.2522**	**0.1091**
0	**0**	**0.04615**	**0.10344**	**0.27016**	**0.4062**	**0.39763**	**0.2934**	**0.22945**	**0.0652**
0.2894	**0.4426**	**0.33144**	**0.45145**	**0.9233**	**1.7194**	**1.9432**	**1.4744**	**1.0635**	**0.4936**
0.4534	**1.4899**	**0.84557**	**0.54043**	**0.90266**	**1.3932**	**1.6188**	**2.4435**	**1.5319**	**0.6683**
0.6789	**0.7570**	**0.76465**	**0.97749**	**1.2574**	**1.7081**	**1.8309**	**2.4021**	**2.3538**	**1.3088**
0.3227	**0.1153**	**0.34051**	**0.1036**	**0.41974**	**0.3746**	**0.65557**	**0.8084**	**1.0096**	**1.0652**

Tabella 6.6. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per durata compresa tra 15 e 60 minuti

0	**0.236**	**0.031**	**0**	**0.026**	**0.014**	**0.025**	**0.035**	**0.030**	**0.02106**
0.23242	**0.23077**	**0.07811**	**0.0300**	**0.0178**	**0.0436**	**0.0438**	**0.0394**	**0.03157**	**0.04848**
0.10848	**0.27217**	**1.2069**	**0.2366**	**0.0838**	**0.0723**	**0.1592**	**0.1742**	**0.105**	**0.13559**
0	**0.0958**	**0.21637**	**2.0811**	**0.5845**	**0.1568**	**0.1094**	**0.1702**	**0.35786**	**0.03273**
0.07181	**0.04877**	**0.06570**	**0.5009**	**1.0508**	**0.5291**	**0.1613**	**0.1611**	**0.26565**	**0.03275**
0.02631	**0.079853**	**0.037952**	**0.09001**	**0.35439**	**0.8125**	**0.31657**	**0.12657**	**0.099642**	**0.048934**
0.20207	**0.34637**	**0.36135**	**0.27165**	**0.46715**	**1.3688**	**0.82955**	**0.64114**	**0.47016**	**0.20814**
0.30027	**0.34257**	**0.43415**	**0.46385**	**0.51245**	**0.6009**	**0.70394**	**1.2174**	**0.68303**	**0.37213**
0.40524	**0.42987**	**0.4102**	**1.5292**	**1.3245**	**0.74178**	**0.80946**	**1.071**	**1.1557**	**0.62774**
0.21748	**0.50907**	**0.40861**	**0.1079**	**0.12598**	**0.281**	**0.27642**	**0.45011**	**0.48422**	**0.36957**

Tabella 6.7. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per durata maggiore a 60 minuti

1.1579	**0.1165**	**0.09162**	**0.02130**	**0.02808**	**0.2949**	**0.1483**	**0.1002**	**0.0538**	**0.05232**
0.11455	**4.5385**	**0.42493**	**0.04570**	**0.0319**	**0.11639**	**0.1306**	**0.1888**	**0.1197**	**0.02808**
0.31389	**1.4806**	**8.1897**	**0.74769**	**0.15373**	**0.13883**	**0.33944**	**0.75106**	**0.68873**	**0.13301**
0.066741	**0.14561**	**0.6836**	**12.1441**	**1.8177**	**0.18021**	**0.25512**	**0.59933**	**0.74229**	**0.088576**
0.075381	**0.08724**	**0.12046**	**1.5578**	**7.4576**	**1.2444**	**0.2541**	**0.21334**	**0.37263**	**0.037164**
0.53019	**0.21282**	**0.072855**	**0.10344**	**0.83339**	**2.0625**	**0.62897**	**0.16395**	**0.20664**	**0.016311**
1.1532	**1.0327**	**0.77027**	**0.63322**	**0.73587**	**2.7197**	**2.3523**	**1.1655**	**0.7168**	**0.32331**
0.85559	**1.6395**	**1.8713**	**1.6332**	**0.67834**	**0.77837**	**1.2797**	**2.0261**	**0.91489**	**0.47671**
0.72097	**1.6298**	**2.6908**	**3.1719**	**1.8579**	**1.5383**	**1.2341**	**1.4346**	**1.7406**	**0.6726**
0.54023	**0.29487**	**0.40085**	**0.29196**	**0.14293**	**0.093668**	**0.42937**	**0.57662**	**0.51882**	**0.86957**

Tabella 6.8. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per luogo "casa"

0.0526	0.1444	0.06427	0.01741	0.0365	0.03026	0.1325	0.12	0.0538	0.053
0.1419	0.1025	0.14991	0.05519	0.0225	0.00701	0.0640	0.23947	0.0537	0.05226
0.2201	0.5223	0.56034	0.44209	0.1732	0.08874	0.2654	0.67582	0.2309	0.23685
0.0545	0.1758	0.40419	0.63964	0.4312	0.14112	0.1044	0.36652	0.2442	0.10764
0.0981	0.0615	0.13572	0.36958	0.6610	0.29544	0.0723	0.14531	0.3034	0.08725
0.0544	0.0128	0.04657	0.081001	0.19787	0.21875	0.11877	0.020682	0.2048	0.021753
1.0301	0.50678	0.60228	0.25932	0.20939	0.51357	0.61364	0.35109	0.48004	0.3151
1.0243	2.0789	1.6838	0.9988	0.46203	0.098191	0.38548	0.98261	0.33366	0.52755
0.72097	0.73153	0.90215	1.0438	1.5131	1.5246	0.82646	0.5232	0.87264	0.58563
0.54719	0.54878	0.71377	0.35482	0.33557	0.12492	0.41846	0.63811	0.45173	0.73913

Tabella 6.9. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per luogo "lavoro"

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0.15517	0	0.042373	0	0.090909	0.023469	0.17202	0.0014303
0	0	0	0	0.033898	0.03125	0.15909	0.021739	0.48113	0
0	0	0.033203	0.029053	0.15254	0.2292	0.071792	0.20323	0.16645	0.015277
0	0	0	0.017938	0.1535	0.21875	0.28552	0.26879	0.12641	0.016311
0	0	0.20629	0.39486	0.20791	1.2346	1.0114	1.2486	0.75977	0.11691
0	0	0.058473	0.059241	0.6462	1.2761	1.3709	2.4696	1.5243	0.18007
0	0	0.67205	2.0559	0.82991	0.94104	1.3081	2.3902	2.1698	0.356
0	0	0.0043103	0	0.058753	0.093668	0.15525	0.21781	0.27461	0.17391

Tabella 6.10. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per luogo "scuola"

0.7368	0.1794	0.03840	0	0	0.25	0.00677	0.01051	0.01061	0
0.17648	4.4615	0.2726	0.01207	0	0.13201	0.04785	0.03675	0.09521	0.00122
0.13158	0.94982	7.5	0.30245	0.011002	0.06410	0.1082	0.25415	0.53387	0.06436
0	0.03846	0.2765	11.2342	1.4607	0.01569	0.04718	0.33267	0.58415	0.00273
0	0	0.0086	1.2519	6.2203	0.12654	0.06593	0.06929	0.2249	0.00220
0.4494	0.24139	0.0336	0.00900	0.084746	0.53125	0.2567	0.00658	0.01127	0.00272
0.0526	0.37836	0.2455	0.11712	0.19096	1.11	0.35227	0.10993	0.07593	0.03226
0.0897	0.31908	0.6332	0.90655	0.22034	0.03125	0.12069	0.31304	0.16296	0.01087
0.1420	1.2962	2.0857	2.4962	1.1213	0.08392	0.13073	0.25553	0.33491	0.05740
0	0.0128	0.1939	0.0090	0.00847	0.0156	0.04284	0.013147	0.044281	0.065217

Tabella 6.11. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per luogo: "altri luoghi" (trasporto,tempo libero)

0.36842	0.25258	0.06858	0.01290	0.0548	0.0292	0.1085	0.0785	0.0802	0.05675
0.24834	0.46154	0.16411	0.02654	0.0667	0.0366	0.1419	0.1765	0.071	0.04261
0.23495	0.5718	2.3793	0.60794	0.17321	0.16266	0.22004	0.45352	0.21908	0.12872
0.040425	0.084569	0.55582	4.955	1.1768	0.32918	0.24683	0.34252	0.1853	0.057334
0.1472	0.18229	0.13572	1.0086	4.5424	1.6798	0.56156	0.26861	0.26438	0.085197
0.052632	0.067033	0.085365	0.18895	1.125	2.5	0.77513	0.30097	0.22787	0.089706
0.84397	1.1227	0.49932	0.61264	1.6263	3.3517	3.4318	1.7002	1.0625	0.55851
0.67003	1.5326	1.1299	0.93339	0.85408	1.4289	1.8668	2.2087	1.2529	0.83464
1.0736	0.97512	0.85592	0.79181	1.3181	1.6963	1.8292	1.9646	2.2594	1.8472
0.5859	0.44747	0.38792	0.18898	0.32766	0.51513	0.74172	1.0095	1.4249	1.6739

Tabella 6.12.Matrice dei contatti medi italiani giornalieri

0.73684	**0.4309**	**0.11373**	**0.00840**	**0.03655**	**0.280**	**0.1415**	**0.1069**	**0.0656**	**0.05742**
0.42371	**4.641**	**0.43292**	**0.06727**	**0.031953**	**0.14603**	**0.10304**	**0.32528**	**0.13527**	**0.069237**
0.38962	**1.5084**	**8.6379**	**0.88514**	**0.17573**	**0.16266**	**0.37738**	**0.95522**	**0.73237**	**0.24057**
0.026316	**0.2143**	**0.80926**	**14.1622**	**2.2029**	**0.21938**	**0.27872**	**0.5477**	**0.95643**	**0.10907**
0.098131	**0.08724**	**0.1377**	**1.888**	**7.8983**	**1.2205**	**0.25032**	**0.30226**	**0.62705**	**0.087401**
0.50388	**0.26703**	**0.085365**	**0.12592**	**0.81741**	**1.4688**	**0.71006**	**0.24693**	**0.26645**	**0.07338**
1.1005	**0.81468**	**0.85636**	**0.69177**	**0.72492**	**3.0703**	**1.8977**	**1.6282**	**0.98561**	**0.39049**
0.91295	**2.8238**	**2.3799**	**1.4925**	**0.96108**	**1.1723**	**1.7877**	**3.113**	**1.4126**	**0.67895**
0.87886	**1.8414**	**2.8612**	**4.087**	**3.1264**	**1.9836**	**1.6969**	**2.2151**	**2.6792**	**0.87403**
0.59286	**0.72702**	**0.72498**	**0.35953**	**0.33614**	**0.42138**	**0.51859**	**0.82124**	**0.6742**	**0.86957**

Tabella 6.13. ITALIA. Indagine Polymod Matrice dei contatti medi italiani per frequenza: "settimanali"

0.42105	**0.0267**	**0.02398**	**0.02191**	**0.0182**	**0.0146**	**0.0169**	**0.0253**	**0.02202**	**0.02361**
0.02631	**0.0512**	**0.0246**	**0.00402**	**0.0310**	**0**	**0.0421**	**0.0248**	**0.02121**	**0.00976**
0.08216	**0.0857**	**0.90517**	**0.24912**	**0.0474**	**0.0246**	**0.1554**	**0.1612**	**0.07981**	**0.09067**
0.06864	**0.0128**	**0.22777**	**1.3874**	**0.4992**	**0.1802**	**0.1169**	**0.2807**	**0.21896**	**0.03**
0.04906	**0.0847**	**0.03716**	**0.42786**	**1.2203**	**0.5619**	**0.2052**	**0.1057**	**0.1196**	**0.05023**
0.02631	**0**	**0.012931**	**0.10344**	**0.37635**	**1.2813**	**0.37438**	**0.10601**	**0.096764**	**0.027195**
0.13158	**0.33354**	**0.35273**	**0.29026**	**0.59427**	**1.6188**	**1.1136**	**0.61194**	**0.40275**	**0.18409**
0.2166	**0.2158**	**0.40177**	**0.76506**	**0.33628**	**0.50329**	**0.67188**	**1.0696**	**0.61173**	**0.27807**
0.29472	**0.28884**	**0.31181**	**0.93565**	**0.5963**	**0.72036**	**0.6934**	**0.95924**	**1.0613**	**0.79351**
0.2438	**0.10256**	**0.27327**	**0.098891**	**0.19321**	**0.15617**	**0.24448**	**0.33634**	**0.61209**	**0.65217**

Tabella 6.14. ITALIA. Indagine Polymod Matrice dei contatti medi per frequenza: occasionali/sporadici

0	**0.05353**	**0.01724**	**0.00840**	**0.0476**	**0.0146**	**0.0530**	**0.0518**	**0.0471**	**0.02361**
0.05263	**0.25641**	**0.08915**	**0.009**	**0.0263**	**0.0296**	**0.0851**	**0.0497**	**0.0480**	**0.00854**
0.05906	**0.31063**	**0.37931**	**0.10267**	**0.1237**	**0.1117**	**0.1119**	**0.1438**	**0.1228**	**0.03518**
0.02631	**0.0287**	**0.09386**	**0.52252**	**0.3125**	**0.1254**	**0.1435**	**0.1394**	**0.1629**	**0.01366**
0.12801	**0.07191**	**0.09692**	**0.2679**	**1.0678**	**0.4383**	**0.2873**	**0.2460**	**0.1667**	**0.04142**
0.026316	**0.054212**	**0.058662**	**0.072032**	**0.29359**	**0.5937**	**0.2788**	**0.2353**	**0.1785**	**0.02991**
0.41259	**0.67355**	**0.25397**	**0.35627**	**0.83206**	**1.2058**	**2.0909**	**1.0314**	**0.87149**	**0.4179**
0.44264	**0.43231**	**0.35844**	**0.37993**	**0.78226**	**1.1175**	**1.1324**	**1.5043**	**1.1038**	**0.51668**
0.63152	**0.65439**	**0.48004**	**0.6963**	**0.83149**	**1.3292**	**1.5004**	**1.7309**	**1.533**	**0.95861**
0.2438	**0.089744**	**0.10603**	**0.045045**	**0.15931**	**0.17179**	**0.55498**	**0.62496**	**0.73945**	**0.67391**

7 Strumenti per l'inferenza su coefficienti di trasmissione: intervalli di confidenza profilo, criteri di 'model selection', bootstrap.

In questo capitolo richiamiamo le nozioni ed i risultati fondamentali relativi ai principali strumenti di inferenza statistica utilizzati nella tesi per sviluppare le procedure inferenziali sui coefficienti di trasmissione. Le prime due sezioni su intervalli di confidenza profilo e criteri di 'model selection', non contiene parti originali e quindi possono essere tranquillamente "saltate" dal lettore non interessato. La terza sezione invece estende le definizioni di stimatori bootstrap al caso di due fonti di incertezza indipendenti, che è il caso del problema di stima di coefficienti di trasmissione

7.1 IC basati sulla verosimiglianza profilo

Un metodo classico di costruzione degli intervalli di confidenza (IC) è basato sulla Normalità asintotica delle stime di Massima verosimiglianza (ML) $\hat{\theta}$ del vettore parametrico θ .
Pertanto, quando, per esempio, μ è un parametro scalare, indicando con $\hat{\mu}$ il corrispondente stimatore ML, l'usuale IC al livello di fiducia (1-α)% è :

$$\hat{\mu} \pm t^{n-1}{}_{1-\alpha/2} SE, \qquad (7.1)$$

ove $t^{n-1}{}_{1-\alpha/2}$ è il quantile (1-α/2) della distribuzione t-student (con n-1 gradi di libertà) o della distribuzione Gaussiana ed *SE* lo standard error del parametro d'interesse.
L'IC così costruito è chiamato anche '*Wald type*'.
Estraendo campioni ripetuti della popolazione, in modo che la teoria asintotica sia applicabile, la copertura dell'IC sarà di circa 1-α ,ovvero vengono lasciate fuori dall'IC tutti i valori campionati che giacciono sulle code simmetriche di probabilità α/2.
E' risaputo inoltre che se la stima di ML del vettore parametrico θ viene ottenuta da un'indagine campionaria di piccola dimensione ,le proprietà di $\hat{\theta}$ possono essere molto differenti da quelle asintotiche. La tecnica nota come '*profile likelihood*'(o anche detta '*metodo del rapporto di verosimiglianza*') può produrre intervalli di confidenza con una migliore copertura. Essa può essere utilizzata nelle analisi statistiche basate sulla verosimiglianza ,anche nel caso in cui il modello abbia più di un parametro . Le stime generate dalla tecnica 'profile likelihood' sono particolarmente utili in modelli non lineari. Tale metodologia può essere usata per una famiglia di modelli di regressione (Venzon e Moolgavkar, 1988).
Nei casi in cui i risultati asintotici non siano applicabili, la copertura reale dell'IC potrebbe essere molto lontana da 1- α . La copertura asimmetrica si verifica quando la distribuzione dello stimatore del parametro d'interesse non è Normale (esempio: distribuzione altamente asimmetrica). Una procedura di costruzione della regione di confidenza che risulti più robusta quando la numerosità campionaria è piccola, può essere basata sulla distribuzione asintotica χ^2 del rapporto di verosimiglianza: l'idea è quella di invertire il test del rapporto di verosimiglianza per il parametro in questione.
Sia θ il vettore parametrico del modello statistico con $dim(\theta)=k>=1$ ed $log(L(\theta))$,la sua funzione di log-verosimiglianza,è quindi definita per valori di θ in uno spazio parametrico k-dimensionale .
Supponiamo adesso di essere interessati alla stima di un singolo elemento (scalare) del vettore dei parametri θ, che indichiamo con β, e che rivesta particolare interesse per i nostri scopi. I parametri

aggiuntivi saranno indicati con δ: L(β, δ) è la funzione di verosimiglianza e $\theta^{ML} = (\beta^{ML}, \delta^{ML})$ è il vettore delle stime di massima verosimiglianza.
La statistica test (G^2) del rapporto di verosimiglianza sotto l'ipotesi nulla $H_0=\beta_0$(dove β_0 è un valore fissato) è uguale alla differenza in 2log(L(.)) tra il modello '*completo*' ed il modello *ridotto* alla condizione $\beta=\beta_0$,

$$G^2 = 2(\log(L(\beta^{ML}, \delta^{ML})) - \log(L(\beta_0, \delta_0^{ML})), \qquad (7.2)$$

dove δ_0^{ML} è la stima di massima verosimiglianza del vettore dei parametri aggiuntivi per il modello 'ridotto'.
Il metodo della verosimiglianza profilo riduce la log-verosimiglianza del vettore parametrico log(L(θ)) ad una funzione di un solo parametro(β) trattando gli altri parametri come '*parametri aggiuntivi'* e massimizzando su di loro.
La statistica test G^2 si può esprimere in termini di funzione di verosimiglianza profilo L_p per il parametro d'interesse β. La funzione di verosimiglianza L_p si ottiene dall'usuale funzione di verosimiglianza L(β, δ) massimizzando il vettore parametrico aggiuntivo δ :

$$L_p(\beta) = \max_{\delta}(L(\beta, \delta)). \qquad (7.3)$$

Dunque utilizzando la log-verosimiglianza profilo $\log(L_p(.))$,G^2 è espressa come :

$$G^2 = 2(\log(L_p(\beta^{ML})) - \log(L_p(\beta_0))). \qquad (7.4)$$

L' intervallo di confidenza per β consiste di quei valori β_0 per cui il test non è significativo al livello di confidenza prestabilito (1- α)% .
Sia α=0.05 il livello di significatività fissato, allora il valore della statistica test G^2, funzione di β_0, non dovrà superare 3.84, il 95° percentile della distribuzione χ^2 con 1 gradi di libertà.
Quindi l'IC 95% basato sulla verosimiglianza profilo consisterà di tutti i valori β_0 per cui

$$\log(L_p(\beta_0) \geq \log(L_p(\beta^{ML}) - (3.84/2) = \log(L(\beta^{ML}, \delta^{ML})) - 1.92. \qquad (7.5)$$

In Fig. 7.1 è rappresentata la verosimiglianza profilo del coefficiente q1 per il modello M2 dei contatti per prossimità e il corrispondente IC di copertura (1-0.05)*100% basato su $L_p(\beta)$.

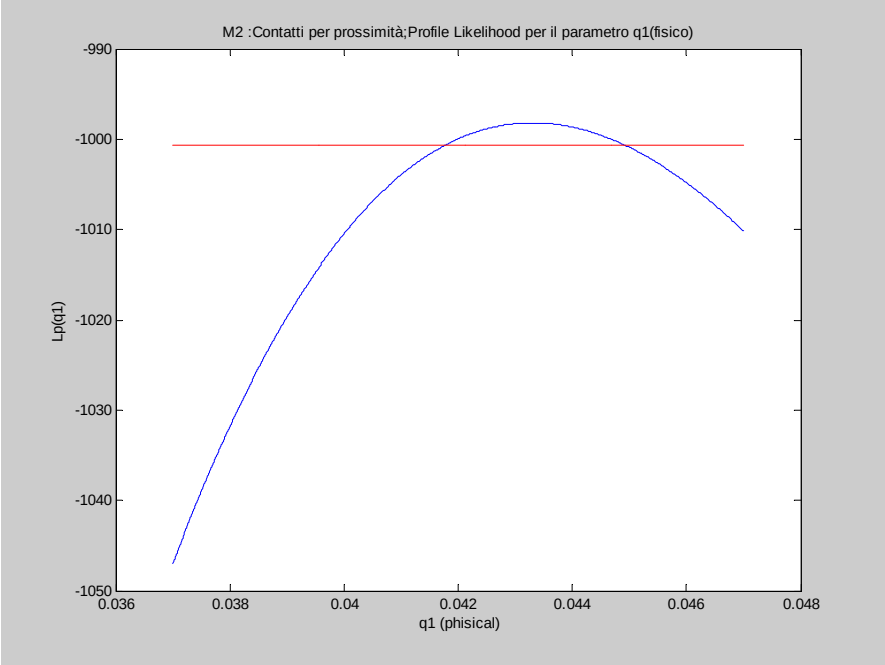


Fig. 7.1:Verosimiglianza profilo e IC 95% per il parametro q2 del modello M2

Per il coefficiente di trasmissione q1 il corrispondente IC 95% basato sulla verosimiglianza profilo è $(\beta_{left}, \beta_{right})=(0.0418;0.0449)$.

La procedura di calcolo dellla funzione di log-verosimiglianza profilo $L_p(\beta)$ e del corrispondente IC con copertura di 100(1-α)% si sviluppa nelle seguenti fasi(per semplicità si considera solo l'estremo inferiore dell'intervallo di confidenza):

1) Si stabilisce un limite inferiore β' ragionevole per stimare l'estremo inferiore dell' IC(i.e. $\beta^*+5SE(\beta^*)$ oppure 0.00001 se il parametro è forzatamente non negativo).
2) Si definisce una griglia di valori compresi tra β' e β^* (i.e. 1000 punti equidistanti),
3) Per ciascun punto della griglia β_i ,si valuta la log-verosimiglianza profilo $\log(L_p(\beta_i))$,massimizzando il vettore dei parametri aggiuntivi della log-verosimiglianza usuale $\log(L(\beta,\delta))$.
4) Si prende come estremo inferiore dell'IC ,β_{left} , il valore più piccolo dei β_i per cui è soddisfatta la disequazione

$$\log(L_p(\beta_i)) \geq \log(L(\beta,\delta)) - 1.92,$$

5) Se necessario ,si ridefinisce o estende la griglia per migliorare l'accuratezza dell'IC.

L'asimmetria percentuale , A%, dell'intervallo di confidenza profilo per il parametro d'interesse può essere calcolata tramite la seguente espressione:

$$A\% = 100\frac{(\beta_{right}-\beta)-(\beta-\beta_{left})}{(\beta_{right}-\beta_{left})} \qquad (7.6)$$

7.2 Criteri di 'model selection' e inferenza multi-modello

7.2.1 L'informazione di Kullback-Leibler

Nel 1951 S. Kullback e RA Leibler pubblicarono un documento ormai famoso che ha quantificato il significato di "informazione " in relazione al concetto di RA Fisher di statistiche sufficienti. Sia f il modello senza parametri che rappresenta la realtà a pieno e g il modello approssimato, una distribuzione di probabilità; l'informazione di Kullback-Leibler I(f,g) è l'informazione perduta quando si usa g per approssimare f ed è definita per funzioni continue dall'integrale

$$I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x/\theta)}\right)dx. \qquad (7.7)$$

Chiaramente il miglior modello perde meno informazioni rispetto agli altri modelli della serie da valutare, questo è equivalente a minimizzare I(f,g) su g. In alternativa, l' informazione KL può essere concettualizzata come una "distanza" tra la realtà e il modello valutato.
Il modello reale è considerato fisso ,solo g varia su uno spazio di modelli indicato con θ.
Chiaramente mentre g varia al variare dell'ampiezza campionaria, la realtà è indipendente dall'numerosità del campione n.
L'informazione KL non può essere usata direttamente come criterio di 'model selection ; essa può essere espressa come

$$I(f,g) = \int f(x)\log(f(x))dx - \int f(x)\log(g(x/\theta))dx$$

$$o \qquad (7.8)$$

$$I(f,g) = E_f[\log(f(x))] - E_f[\log(g(x/\theta))],$$

La quantità E(log(f(x))) è costante e non dipende né dai dati né dal modello:

$$I(f,g) = C - E_f[\log(g(x/\theta))], \qquad (7.9)$$

con

$$C = \int f(x)\log(f(x))dx. \qquad (7.10)$$

7.2.2 Il criterio di informazione di Akaike

Akaike (1973, 1974, 1985, 1994) ha mostrato che il punto critico per ottenere un rigoroso criterio di selezione del modello sulla base delle informazioni KL è stato quello di stimare

$$E_y E_x[\log(g(x/\hat{\theta}(y)))] \qquad (7.11)$$

dove la parte interna è proprio E_f [log(g(x|θ))], con θ sostituito dal suo stimatore di MV basato sul modello g assunto e sui dati y. Anche se solo y rappresenta i dati, si deve sottolineare che sia x che y sono campioni casuali indipendenti estratti dalla stessa distribuzione.
Entrambe i valori attesi sono calcolati rispetto al modello vero f .
Akaike ha trovato una relazione formale tra l'informazione KL e la teoria della probabilità: il valore che massimizza la log-verosimiglianza è uno stimatore distorto di

$$E_y E_x[\log(g(x/\hat{\theta}(y)))]. \qquad (7.12)$$

Tale distorsione è approssimativamente uguale a K, il numero di parametri da stimare nel modello approssimato g: questo è un risultato asintotico di fondamentale importanza. Quindi uno stimatore approssimatamente corretto di (7.14) per campioni grandi è $\log(L(\hat{\theta}/dati)) - K$.
Questo risultato equivale ad affermare che

$$\log(L(\hat{\theta}/dati)) - K = C - \hat{E}_{\hat{\theta}}[I(f,\hat{g})], \qquad (7.13)$$

dove

$$\hat{g} = g(.|\hat{\theta}). \qquad (7.14)$$

Questa scoperta rende possibile la combinazione di stime (cioè, di massima verosimiglianza o dei minimi quadrati) e la selezione del modello ottimo nell'ambito di un quadro unificato.
Akaike ha trovato un stimatore dell'informazione KL attesa relativa, basato sulla funzione di log-verosimiglianza massimizzata, corretto per la distorsione asintotica,

$$relative \quad \hat{E}(K-L) = \log(L(\hat{\theta}/dati)) - K. \qquad (7.15)$$

K è il termine asintotico di correzione della distorsione e non è in alcun modo arbitrario. Il criterio di informazione di AKAIKE (AIC) è quindi definito (per ragioni storiche) come:

$$AIC = -2\log(L(\hat{\theta}/dati)) + 2K. \qquad (7.16)$$

Assumendo una serie di modelli candidati, l'AIC è calcolato per ciascuno di essi. Utilizzando l'AIC i modelli possono essere facilmente classificati dal migliore al peggiore sulla base dei dati empirici. Questo è un concetto semplice ed interessante, basato su saldi fondamenti teorici (cioè, entropia, informazione KL, verosimiglianza statistica) (Burnham and Anderson(2002)).
.

7.2.3 Le differenze di Akaike

I singoli valori AIC non sono interpretabili in quanto contengono costanti arbitrarie e sono molto influenzati dalle dimensioni del campione. Si definisce la differenza di Akaike per il modello g_i considerato tramite la seguente formula:

$$\Delta_i = AIC_i - AIC_{\min}, \qquad (7.17)$$

dove AIC_{min} è il minimo dei diversi valori AIC .

Il modello migliore ha differenza Akaike pari a 0, mentre per il resto dei modelli le differenze Akaike assumono valori positivi. La costante $C = \int f(x)\log(f(x))dx$ viene eliminata e non influenza la misura della differenza di Akaike.
Quindi, Δ_i è la perdita di informazioni attesa se si utilizza il modello g_i rispetto al modello migliore. I Δ_i consentono un'interpretazione efficace senza che costanti di scala e dimensione del campione ne influenzino la misura. Sono facili da interpretare e consentono un confronto veloce della "forza dell'evidenza" tra i modelli candidati. Più grande è la differenza di Akaike meno plausibile è il modello considerato; è inoltre possibile tramite semplici regole di soglia capire quali tra i modelli sono abbastanza plausibili e quali non lo sono affatto. Tali regole sono usate anche in ambito Bayesiano(Raftery 1996).

7.2.4 I pesi Akaike

La semplice trasformazione $\exp(-\Delta_i/2)$ fornisce la verosimiglianza del modello: $L(g_i|\text{ dati})$.
Poiché l'AIC viene moltiplicato per -2, $L(g_i|\text{ dati})=\exp(\Delta_i)$; questa è la funzione di verosimiglianza sull'insieme dei modelli nel senso che $L(\theta|\text{dati},g_i)$ è la verosimiglianza sullo spazio parametrico (per il modello g_i) del parametro θ, condizionatamente ai dati x e al modello g_i. La verosimiglianza relativa del modello i rispetto al modello j è data dal rapporto ***L(g_i\| dati)/ L(g_j\| dati)***; questa misura è il rapporto dell'evidenza (***Evidence Ratio***(ER)) e non dipende da nessuno degli altri modelli considerati. Assumendo che g_i sia migliore di g_j, l''Evidence ratio' sarà un valore grande e di conseguenza il modello g_j sarà un modello povero relativamente al modello g_i , basandoci sui dati.
Conviene normalizzare le verosimiglianze del modello in modo che la loro somma faccia 1 per trattarle come funzioni di probabilità; i pesi Akaike w_i e il rapporto dell'evidenza ER, quando si considerano R possibili modelli, sono quindi dati dalle seguenti espressioni :

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{i=1}^{R}\exp(-\Delta_r/2)}. \qquad (7.18)$$

$$ER = \frac{w_{\min}}{w_r}$$

Essi pesano l'evidenza in favore del modello $g_i(.|\theta)$: il rapporto w_i/w_j è identico al rapporto $L(g_i|\text{dati})/L(g_j|\text{dati})$, cosicché siano invarianti per l'insieme dei modelli considerati(es: g_i, g_j), ma i valori w dipendono dall'insieme completo dei modelli poiché la loro somma è uno.
I pesi di Akaike possono quindi essere interpretati come la probabilità che il modello i sia il migliore in termini di distanza di K-L.

7.2.5 Stimatori non condizionati e inferenza multi-modello

Quando si usano criteri di 'model selection', una componente della varianza deriva dall'incertezza sul modello da scegliere: questa componente può essere considerata nello stimatore della varianza, generando stime incondizionate dai criteri di scelta .Un semplice stimatore della varianza incondizionata per lo stimatore di MV per il modello selezionato è:

$$\hat{\text{var}}(\hat{\theta}) = \left[\sum_{i=1}^{R} w_i\left[\text{var}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2\right]^{1/2}\right]^2$$

dove (7.19)

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i\hat{\theta}_i$$

rappresenta una forma di '***model averaging***'.
Il primo termine dello stimatore della varianza è la varianza campionaria condizionata dalla scelta del modello g_i, mentre il secondo termine è la componente di varianza che deriva dall'incertezza del criterio di 'model selection'.
Queste componenti sono moltiplicate per i pesi di Akaike per valutare l'evidenza relativa dell'i-esimo modello. Attraverso il bootstrap o repliche Monte Carlo è possibile ottenere anche intervalli di confidenza per tale parametro: questa metodologia assicura una copertura effettiva molto vicina a quella nominale.

7.3 Inferenza bootstrap: introduzione

I dati utilizzati nel modello statistico di stima dei parametri di trasmissione sono i dati di contatto ed i dati sierologici. Anche se le due fonti sono indipendenti, essendo il modello altamente non lineare l'incertezza totale di stima dipende in maniera complessa dalle incertezze di fondo delle due fonti. Le procedure inferenziali sui parametri di trasmissione q (e susseguentemente sugli R_0) possono essere convenientemente sviluppate tramite metodi di ricampionamento delle due fonti dei dati.
Il ricampionamento bootstrap di tipo 'standard' è casuale con reimissione: ciò che rimane inalterato per ogni campione bootstrap è la numerosità campionaria pari all'ampiezza campionaria dell'indagine. In questa tesi si utilizza il Bootstrap non parametrico delle 2 fonti di dati per costruire intervalli di confidenza dei parametri d'interesse.(Efron e Tibshirani, 1993).
La strategia adottata è il ricampionamento diretto degli individui che hanno partecipato alle 2 indagini campionarie.
In generale le motivazioni teoriche dell'utilizzo di tecniche di ricampionamento vanno ricercate nella constatazione che solo in casi semplicissimi siamo in grado di fornire delle indicazioni dello standard error delle stime dal solo campione osservato: questo caso è quello della media della popolazione normale (stimiamo lo standard error mediante la radice dello stimatore corretto della varianza) oppure della media di una popolazione qualunque (ma in tal caso dobbiamo possedere molti dati per invocare il Teorema del Limite Centrale). Il bootstrap si propone come una strategia 'naturale' per fare inferenza anche su parametri per cui non possediamo un semplice stimatore dell'errore standard (es: i-esimo percentile). Il metodo del bootstrap non-parametrico è un'applicazione del principio del plug-in. Il principio del plug-in è definito considerando che un ovvio modo di stimare alcuni aspetti della distribuzione della popolazione è usare i corrispondenti aspetti della distribuzione di densità empirica(Efron e Tibshirani, 1993).
Non-parametrico significa che tale metodologia utilizza solo l'informazione contenuta nei dati osservati, mentre non si impone nessuna conoscenza a priori sulla distribuzione di densità della popolazione f. Originariamente Bradley Efron nel 1979 inventò il bootstrap[2] per calcolare lo standard error di uno stimatore arbitrario. Nelle seguenti sezioni vengono definite misure dell'accuratezza (se,cv,distorsione) stimabili attraverso procedure di ricampionamento bootstrap in presenza di due fonti d'incertezza indipendenti.
Inoltre si generalizzano alle due fonti d'incertezza le procedure di costruzione degli intervalli di confidenza bootstrap , esposte nel classico " An Introduction to the Bootstrap " scritto da Efron e Tibshirani (1993).

[2] Il termine 'bootstrap' deriva dalla frase " to pull oneself up by one's bootstrap " (riuscire con le proprie forze) tratto da 'Le avventure del Barone Munchausen' di Rudolf Erich Raspe. Il barone era caduto sul fondo di un lago profondo:quando realizzò che ogni speranza di sopravvivere era ormai perduta, pensò di salvare se stesso con le proprie forze .

7.4 Standard error (SE) Bootstrap

La strategia adottata è il ricampionamento diretto con reimissione degli individui che hanno partecipato alle 2 indagini campionarie. Pertanto, indicato con θ il vettore dei parametri incogniti ed s lo stimatore desiderato, la procedura si sviluppa nei seguenti passi:

(A) si selezionano (casualmente con reimmissione) *P* campioni bootstrap indipendenti x_1*, x_2*,…x_P*, dai dati sierologici (1° fonte di incertezza) e *B* campioni bootstrap indipendenti y_1*, y_2*,… y_B*, dai dati di contatto (2° fonte di incertezza).

(B) Si calcola la replica bootstrap $\hat{\theta}^*(b,p)$, attraverso il principio del plug-in, per ognuna delle *B*P* coppie congiunte (x_b*, y_p*,) di campioni bootstrap :

$$\hat{\theta}^*(b,p) = s(x^{*b}, y^{*p}), \quad b = 1....B, \quad p = 1.....P. \tag{7.20}$$

Le repliche bootstrap $\hat{\theta}^*(b,p)$ sono le rispettive stime plug-in della statistica s(.) calcolate per la coppia di campioni bootstrap x^{*b} e y^{*p} .

(C) Si stima lo standard error dello stimatore del parametro d'interesse tramite la deviazione standard campionaria delle B*P repliche :

$$SE_{\hat{F}_1,\hat{F}_2}(\hat{\theta}) = \hat{SE}_{B,P}\left\{\sum_{b=1}^{B}\sum_{p=1}^{P}\frac{\left[\hat{\theta}^*(b,p) - \hat{\theta}^*(.)\right]^2}{(B-1)(P-1)}\right\}^{1/2}, \tag{7.21}$$

dove

$\hat{\theta}^*(.) = \sum_{b=1}^{B}\sum_{p=1}^{P}\frac{\hat{\theta}^*(b,p)}{B*P}$ è il valor medio delle repliche bootstrap $\hat{\theta}^*(b,p)$.

Lo standard error bootstrap è quindi la stima plug-in ottenuta dalla deviazione standard delle repliche bootstrap .

7.5 Standard error ideale

La stima bootstrap ideale (Efron e Tibshirani, 1993) dello standard error è il valore dello standard error bootstrap calcolato per un numero infinito di repliche della procedura di stima del parametro d'interesse:

$$\lim_{\substack{B\to\infty \\ P\to\infty}}(\hat{SE}_{B,P}) = SE_{\hat{F}_1,\hat{F}_2} = SE_{\hat{F}_1,\hat{F}_2}(\hat{\theta}^*), \tag{7.22}$$

dove $\hat{F}_1, \hat{F}_2$ sono le distribuzioni empiriche delle due fonti dei dati. Empiricamente si possono utilizzare vari criteri per valutare la stabilità dello standard error bootstrap. Un criterio molto comune è basato sul coefficiente di variazione CV, fissando una soglia sul CV sotto la quale lo standard error è ritenuto "stabile"(Efron e Tibshirani, 1993).

$$\hat{CV}_{\hat{F}_1,\hat{F}_2} = \frac{\hat{SE}_{\hat{F}_1,\hat{F}_2}}{\hat{\theta}^*(.)} \tag{7.23}$$

Il coefficiente di variazione bootstrap è il rapporto tra standard error bootstrap e valore atteso bootstrap.

7.6 Distorsione bootstrap

Si vuole stimare il valore vero $\theta = t(F_1, F_2)$ del parametro d'interesse tramite la sua stima non parametrica :

$$\hat{\theta} = t(\hat{F}_1, \hat{F}_2). \qquad (7.24)$$

La distorsione "reale" della stima di uno stimatore è la differenza tra il valor atteso della stimatore del parametro e il valore vero del parametro. Nell'approccio bootstrap la distorsione "reale" (sconosciuta) viene stimata mediante la distorsione bootstrap che si ottiene sostituendo alla distribuzione reale del parametro(sconosciuta) la distribuzione empirica dei dati secondo il principio del plug-in:

$$\begin{aligned}
&bias_{F_1,F_2} = bias_{F_1,F_2}(\hat{\theta}, \theta) = E_{F_1,F_2}\left(\hat{\theta}\right) - \theta \\
&bias_{\hat{F}_1,\hat{F}_2} = bias_{\hat{F}_1,\hat{F}_2}(\hat{\theta}^*, \hat{\theta}) = E_{\hat{F}_1,\hat{F}_2}\left(\hat{\theta}^*\right) - \hat{\theta}, \\
&E_{\hat{F}_1,\hat{F}_2}\left(\hat{\theta}^*\right) \cong \hat{\theta}^*(.) = \sum_{b=1}^{B}\sum_{p=1}^{P}\frac{\hat{\theta}^*(b,p)}{B*P} \\
&\hat{bias}_{B,P} = \hat{\theta}^*(.) - \hat{\theta}.
\end{aligned} \qquad (7.25)$$

La distorsione bootstrap del parametro è quindi data dalla differenza tra il valor atteso delle repliche bootstrap e la stima puntuale del parametro $\hat{\theta}$=s(x,y), ottenuta dai campioni originali.
Ma quando la distorsione della stima del parametro è trascurabile?
Per rispondere a tale domanda è necessario indagare sulla relazione tra standard error e bias .
Affinchè la distorsione sia trascurabile, è necessario appurare che, al livello di repliche scelto, il rapporto tra la stima bootstrap della distorsione e la stima bootstrap dello standard error sia inferiore ad una soglia prefissata; una regola applicativa comunemente usata (Efron e Tibshirani, 1993) è assicurarsi che tale rapporto sia inferiore a 0.25 al fine di ottenere IC affidabili:

$$\begin{aligned}
&\sqrt{E_{F_1,F_2}[(\hat{\theta}-\theta)]^2} = \sqrt{SE_{F_1,F_2}(\hat{\theta})^2 + bias_{F_1,F_2}(\hat{\theta},\theta)^2} = SE_{F_1,F_2}(\hat{\theta})\sqrt{1+\left(\frac{bias_{F_1,F_2}}{se_{F_1,F_2}}\right)^2} \\
&SE_F(\hat{\theta})\sqrt{1+\left(\frac{bias_{F_1,F_2}}{se_{F_1,F_2}}\right)^2} \cong SE_F(\hat{\theta})\left[1+\frac{1}{2}\left(\frac{bias_{F_1,F_2}}{se_{F_1,F_2}}\right)^2\right] \\
&\left|\frac{bias_{\hat{F}_1,\hat{F}_2}}{SE_{\hat{F}_1,\hat{F}_2}}\right| < 0.25
\end{aligned} \qquad (7.26)$$

7.7 Intervalli di confidenza bootstrap

Riportiamo di seguito brevi nozioni sui principali stimatori bootstrap utilizzati in letteratura ed applicati in questa tesi.

7.7.1 Intervalli di confidenza basati sull'assunzione di Normalità e sullo SE ideale

La distribuzione delle repliche bootstrap è l'informazione fondamentale per la costruzione di qualsiasi tipo di IC , se non si vogliono fare assunzioni a priori sulla distribuzione del parametro .

Chiaramente se la numerosità del campione di partenza è grande può essere utile confrontare la distribuzione delle repliche bootstrap standardizzata con la distribuzione Normale Standard
Attraverso un confronto dei quantili delle 2 distribuzioni è possibile sottoporre a verifica l'ipotesi che tali distribuzioni siano uguali.
Se si assume che la distribuzione delle repliche è distribuita normalmente allora:

$$\hat{\theta}^* \approx N(\hat{\theta}, \hat{SE}_{B,P}) \qquad (7.27)$$

Che conduce all'intervallo di confidenza $IC_{normal}100(1-\alpha)\% = \left\lfloor \hat{\theta}_{low}, \hat{\theta}_{up} \right\rfloor$ ove:

$$\hat{\theta}_{low} = \hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{SE}_{B,P\to\infty}$$

è l'estremo inferiore e

$$\hat{\theta}_{up} = \hat{\theta} + z_{1-\frac{\alpha}{2}} \hat{SE}_{B,P\to\infty}$$

l'estremo superiore dell'IC bootstrap Normale.

7.7.2 Intervalli di confidenza Bootstrap-T (studentizzati)

Senza fare assunzioni sulla normalità della distribuzione delle repliche bootstrap, si può ottenere l'IC bootstrap-T stimando la distribuzione di $Z = \frac{\hat{\theta} - \theta}{SE}$ direttamente dai dati, tramite il principio del plug-in .
La procedura di calcolo dell' IC bootstrap-T generalizzata per considerare due fonti d'incertezza è di seguito presentata:

1. Si generano B*P campioni bootstrap dalle due fonti (livello 1) e si calcolano le corrispondenti B*P repliche bootstrap del parametro d'interesse e lo standard error bootstrap associato alle B*P repliche.
2. Ognuno dei B campioni bootstrap della prima fonte dei dati viene nuovamente ricampionato B_2 volte generando campioni bootstrap di livello 2, e ognuno dei P campioni bootstrap della seconda fonte dei dati viene ricampionato generano P_2 campioni bootstrap di livello 2, ottenendo complessivamente $B*P*B_2*P_2$ repliche congiunte dei due livelli .
3. Per ognuna delle B_2*P_2 repliche del parametro d'interesse viene calcolato lo standard error condizionato ai particolari campioni bootstrap di 1° livello: complessivamente ottengo B*P valori dello standard error condizionato. Tale standard error è detto anche di secondo livello dato che per ottenerlo bisogna ricampionare un campione bootstrap di primo livello.
4. La distribuzione Z*(.) è determinata tramite la stima dello standard error della particolare replica bootstrap del parametro di interesse (SE di 2° livello):

$$Z^*(b,p) = \frac{(\hat{\theta}^*(b,p) - \hat{\theta})}{\hat{SE}^*(b,p)} \qquad (7.28)$$

è quindi la stima plug-in di Z ,che si ottiene utilizzando l'informazione empirica dei ricampionamenti bootstrap di primo livello $\hat{\theta}*(b,p) = s(x^*(b,p))$ e di secondo livello $\hat{\theta}^{**}(b_2, p_2) = s(x_b^{**}, y_p^{**})$,dove x_b^{**} e y_p^{**} sono rispettivamente i ricampionamenti bootstrap del b-esimo e del p-esimo campione bootstrap di livello 1.
Per ognuna delle B*P replica bootstrap del parametro al livello 1 si ottiene uno standard error bootstrap di livello 2 dato dalla seguente formula:

$$\hat{SE}^*(b,p)=\sqrt{\sum_{aux=1}^{B_2*P_2}\frac{\left(\hat{\theta}^{**}(b_2,p_2)-\hat{\theta}^*(b,p)\right)^2}{(B_2-1)(P_2-1)}} \qquad (7.29)$$

5. Il quantile standardizzato α è determinato dalla stima di $\hat{t}^{(\alpha)}$ in modo che :

$$\frac{\#\{Z^*(b,p)\le\hat{t}(\alpha)\}}{BP}=\alpha \qquad (7.30)$$

6. L'IC bootstrap-t (1- α)% è dato dai seguenti valori limite:

$$[\hat{\theta}_{low},\hat{\theta}_{up}]=\left[\hat{\theta}-\hat{t}^{(1-\frac{\alpha}{2})}\hat{se};\hat{\theta}-\hat{t}^{(\frac{\alpha}{2})}\hat{se}\right], \qquad (7.31)$$

dove $\hat{SE}$ è lo standard error bootstrap del parametro d'interesse .

Come mostrato in letteratura l'intervallo studentizzato non è simmetrico, è inoltre erratico (risente degli outliers). L'accuratezza dell'IC bootstrap-t è del secondo ordine: ovvero, affermando che la copertura effettiva dell'IC 'studentized' sia (1-α), si commette un errore più piccolo rispetto a quando l' IC ha accuratezza del primo ordine.
Se l'accuratezza è del secondo ordine allora:

$$\Pr(\theta<\hat{\theta}_{low})\cong\frac{\alpha}{2}+\frac{k_{low}}{n}\quad\cap\quad\Pr(\theta>\hat{\theta}_{up})=\frac{\alpha}{2}+\frac{k_{up}}{n}, \qquad (7.32)$$

mentre per IC con accuratezza al primo ordine vale la sequenza condizione(esempio:IC bootstrap Normale)

$$\Pr(\theta<\hat{\theta}_{low})\cong\frac{\alpha}{2}+\frac{k_{low}}{\sqrt{n}}\quad\cap\quad\Pr(\theta>\hat{\theta}_{up})=\frac{\alpha}{2}+\frac{k_{up}}{\sqrt{n}}. \qquad (7.33)$$

Nel prossimo paragrafo verranno analizzate altre due proprietà tra loro collegate dell'IC 'studentized ': l'IC studentizzato è stabile quando la varianza dello stimatore è costante. Sia Φ=g(θ) una trasformazione del parametro d'interesse: solo nel caso in cui tale trasformazione rende costante la varianza di θ, l'intervallo di confidenza costruito su Φ può essere trasformato correttamente , secondo la funzione inversa $g^{-1}(\Phi)$, in un intervallo per θ.
Quando la trasformazione utilizzata non stabilizza la varianza del parametro θ, non è lecito trasformare l'intervallo di confidenza per Φ, tramite la trasformazione inversa: in tal senso si può affermare che l'IC 'studentized' non è '*trasformation respecting*'.

7.7.3 Intervallo bootstrap-t a varianza stabilizzata

L'IC bootstrap studentizzato ha l'importante caratteristica di essere affidabile se la varianza dei parametri è costante. Sia X una variabile aleatoria con media θ e varianza SE(θ), allora la derivata della trasformazione g(X)

$$g'(X)=\frac{1}{SE(x)} \qquad (7.34)$$

rende costante la varianza di g(X).
Quindi, in un intorno opportuno del valore del parametro, applicando una trasformazione g(x) a X ed espandendo g(x) in serie di Taylor si ottiene la seguente approssimazione:

$$Var(g(X))=Var\left(g(\theta)+g'(\theta)(X-\theta)+...\right)\approx$$
$$\approx Var(g(\theta))+Var\left(g'(\theta)(X-\theta)\right)+2\operatorname{cov}\left(g(\theta)g'(\theta)(X-\theta)\right)= \qquad (7.35)$$
$$=g'(\theta)^2\operatorname{var}(X),$$

Pertanto ne segue che la trasformazione

$$g(x) = \int^{x} \frac{1}{SE(u)}\, du,$$

$$dove \qquad (7.36)$$

$$SE(u) = SE(X / \theta = u),$$

ha varianza costante $\text{var}(g(x)) \approx \cos\tan te$.

Per il nostro problema X è $\hat{\theta}$ ed ad ogni u si desidera conoscere lo standard error s(u) per applicare la trasformazione(7.35).n Naturalmente $SE(\hat{\theta} / \theta = u)$ è solitamente sconosciuto, ma si può utilizzare il bootstrap per stimarlo. Quindi si calcola l'intervallo bootstrap-t per la trasformazione $\Phi=g(\theta)$ e solo successivamente al calcolo dell'IC si utilizza la trasformazione inversa $g^{-1}(.)$ per ottenere l'intervallo per il parametro d'interesse θ (Tibshirani, 1988)
La procedura necessaria ad ottenere IC bootstrap-t a varianza stabilizzata nel caso di 2 fonti d'incertezza è esposta di seguito:

1. Si generano B campioni bootstrap dalla prima fonte dei dati campionari e P campioni bootstrap dalla seconda fonte dei dati campionari, ottenendo B*P repliche bootstrap della statistica d'interesse. Si calcola, poi, secondo la formula(7.23), lo standard error bootstrap del parametro d'interesse θ.
2. Per ognuno dei B(P) campioni bootstrap di livello 1 si ottengono $B_2(P_2)$ campioni bootstrap di secondo livello ed un totale di $B*B_2*P*P_2$ repliche bootstrap di secondo livello. Per ogni coppia di campioni bootstrap (al livello 1) (x*(b),x*(p)) con b=1…B e p=1…P si ottengono quindi B_2*P_2 repliche di secondo livello attraverso le quali si calcolano B*P stime dello standard error (al livello 2) $\hat{SE}(\hat{\theta}*(b,p))$.
3. Si adatta una curva ai punti $[\hat{\theta}*(b,p), \hat{SE}(\hat{\theta}*(b,p))]$ per ottenere una funzione continua('smooth') di $SE(u) = SE(\hat{\theta} / \theta = u)$ tramite stima adattiva (es:spline cubiche).
4. Si utilizza la trasformazione $g(\hat{\theta})$ che stabilizza la varianza e si risolve numericamente l'integrale della formula (7.35).
5. Si utilizzano B_3 e P_3 nuovi campioni bootstrap di livello 1 per calcolare l'intervallo di confidenza bootstrap-t per la trasformazione $\Phi=g(\theta)$

(Si noti che poiché la varianza della trasformazione di Φ è costante non bisogna calcolare il denominatore della statistica bootstrap-t $(g(\hat{\theta}^*) - g(\hat{\theta}) / \hat{SE}^*)$.

6. Si utilizza la trasformazione inversa $g^{-1}(.)$ per ottenere l'intervallo IC-bootstrap di $\hat{\theta}$.

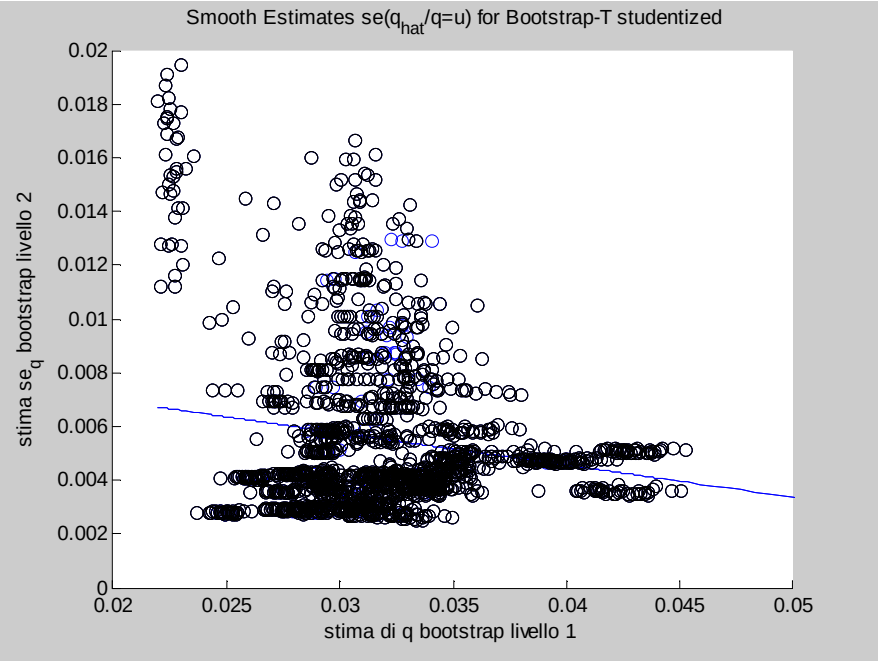


Fig. 7.2: Stima smooth dello SE bootstrap (livello 2) per il coefficiente di trasmissione q (Modello M1)tramite spline cubiche

In Fig. 7.2 sono mostrate(in nero) le stime dello standard error bootstrap(livello 2) condizionatamente ai campione bootstrap(livello 1) ricampionati, l'andamento(in blu) mostra la stima adattiva ottenuta tramite 'spline cubica' con parametro di smooth pari a 0.3.
La Fig. 7.3 invece descrive la trasformazione della stima smooth dello standard error condizionato.

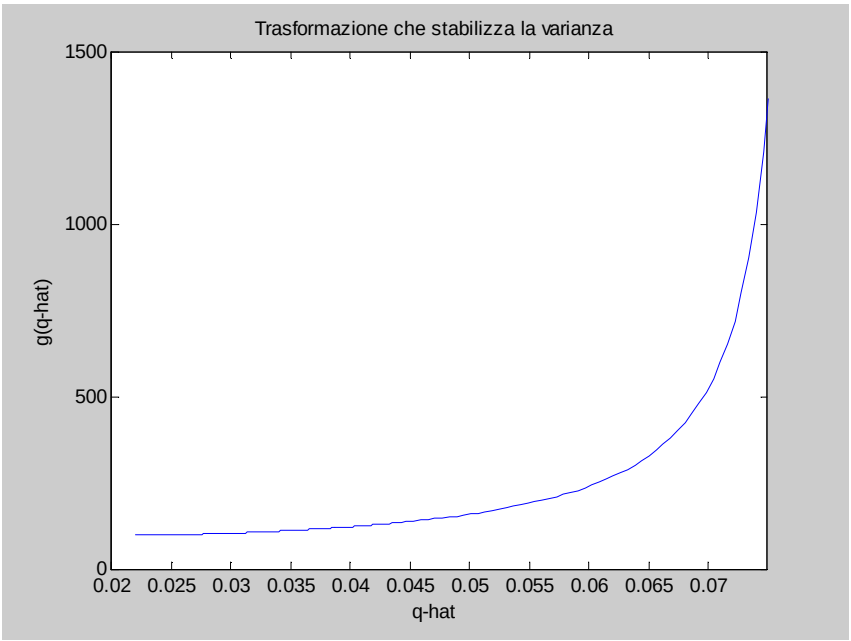


Fig. 7.3:Trasformazione del parametro q che stabilizza la varianza(Modello M1)

7.7.4 Intervalli di confidenza percentili di Efron

L'idea di Efron è quella di utilizzare la distribuzione delle repliche bootstrap del parametro d'interesse per ottenere IC senza alcuna assunzione sulla distribuzione. Sia $\hat{G}$ la funzione di distribuzione cumulativa delle repliche bootstrap del paramentro d'interesse θ, l'IC 100(1-α)% per tale parametro è definito tramite il percentile α/2 e (1- α/2) di $\hat{G}$

$$[\hat{\theta}_{\%,low}, \hat{\theta}_{\%,up}] = [G^{-1}(\alpha/2), G^{-1}(1-\alpha/2)] \quad (7.37)$$

Poichè per definizione è $G^{-1}(\alpha/2) = \hat{\theta}^{*(\frac{\alpha}{2})}{}_{B,P}$, il 100(1-α/2) percentile della distribuzione bootstrap, allora vale la seguente identità:

$$[\hat{\theta}_{\%,low}, \hat{\theta}_{\%,up}] = [\hat{\theta}^{*(\frac{\alpha}{2})}{}_{B,P}, \hat{\theta}^{*(1-\frac{\alpha}{2})}{}_{B,P}] \quad (7.38)$$

L'Efron percentile IC non è(necessariamente) simmetrico ed ha un'accuratezza del primo ordine ovvero si approssima la probabilità di copertura ad un errore dell'ordine di grandezza pari a $n^{-1/2}$ (dove n è la numerosità campionaria).
L'IC percentile è inoltre 'trasformation respecting ', ovvero l'IC percentile di qualsiasi trasformazione monotona del parametro Φ=g(θ) è semplicemente l'IC percentile di θ calcolato da g(θ).

7.7.5 Intervalli di confidenza BCa (Bias Corrected and Accellerated)

L'IC costruito tramite il metodo BCa, dove BCa indica che tale metodo implica una correzione della distorsione e la definizione dell'accellerazione (Efron e Tibshirani, 1993), si ottiene direttamente, come l'IC percentile, dalla distribuzione delle repliche bootstrap del parametro d'interesse.
La scelta degli estremi dell'IC dipende da 2 numeri $\hat{a}$ e $\hat{z}_0$ chiamate accellerazione e correzione della distorsione . L'IC 100(1-α)% BCa è dato da :

$$(\hat{\theta}_{low}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{(\alpha_2)}), \quad (7.39)$$

dove

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\frac{\alpha}{2})}}{1 - \hat{a}(\hat{z}_0 + z^{(\frac{\alpha}{2})})}\right)$$
$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\frac{\alpha}{2})}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\frac{\alpha}{2})})}\right). \quad (7.40)$$

Φ(.) è la distribuzione cumulativa Normale standard e z(α/2) è il percentile 100α-esimo della distribuzione Normale standard. Si noti che se $\hat{a} = 0$ e $\hat{z}_0 = 0$, l'intervallo di confidenza BCa coincide con l'intervallo di confidenza Normale.
Il valore della correzione della distorsione si calcola direttamente dalla proporzione delle repliche 'bootstrap' minori della stima del parametro d'interesse :

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b,p) < \hat{\theta}\}}{(B*P)}\right). \quad (7.41)$$

Per calcolare l'accelerazione si usano i valori 'jackknife' (Efron e Tibshirani, 1993) della statistica $\hat{\theta} = s(x,y)$ .Siano x(i) e y(k) le 2 fonti dei dati originali a cui è stata sottratta l'i-esima e la k-esima osservazione rispettivamente, ovvero uno dei possibili campioni jackknife e sia $\hat{\theta}(i,k) = s(x(i), y(k))$ allora:

$$\hat{a} = \frac{\sum_{i=1}^{n_x}\sum_{k=1}^{n_y}\left(\hat{\theta}(.) - \hat{\theta}(i,k)\right)^3}{6\left\{\sum_{i=1}^{n_x}\sum_{k=1}^{n_y}\left(\hat{\theta}(.) - \hat{\theta}(i,k)\right)^2\right\}^{\frac{3}{2}}}, \qquad (7.42)$$

dove

$$\hat{\theta}(.) = \frac{\sum_{i=1}^{n_x}\sum_{k=1}^{n_y}\left(\hat{\theta}(i,k)\right)}{n_x n_y}. \qquad (7.43)$$

L' accelerazione esprime il tasso di cambio dello standard error di $\hat{\theta}$ rispetto al valore vero del parametro θ. L'approssimazione Normale Standard $\hat{\theta} \approx N(\theta, se^2)$ assume che lo standard error di $\hat{\theta}$ sia costante al variare di θ.
La correzione della distorsione è il 'bias' mediano della replica bootstrap, ovvero la differenza tra la mediana di $\hat{\theta}^*$ (distribuzioni delle repliche bootstrap) e la stima puntuale del parametro d'interesse $\hat{\theta}$.
L'IC BCa continua ad essere valido anche quando viene utilizzata una trasformazione monotona del parametro come per l'IC percentile .
Inoltre è asimmetrico ed ha un'accuratezza del secondo ordine:

$$\Pr(\theta < \hat{\theta}_{low}) \cong \frac{\alpha}{2} + \frac{k_{low}}{n} \quad \cap \quad \Pr(\theta > \hat{\theta}_{up}) = \frac{\alpha}{2} + \frac{k_{up}}{n} \qquad (7.44)$$

L'accuratezza del primo ordine (es: IC normale e percentile) implica:

$$\Pr(\theta < \hat{\theta}_{low}) \cong \frac{\alpha}{2} + \frac{k_{low}}{\sqrt{n}} \quad \cap \quad \Pr(\theta > \hat{\theta}_{up}) = \frac{\alpha}{2} + \frac{k_{up}}{\sqrt{n}}. \qquad (7.45)$$

7.7.6 Standard error e distorsione Jack-knife

La stima dello standard error e della distorsione bootstrap può essere confrontata con quella jackknife(one-deleted). 'One-deleted' significa che i ricampionamenti dai dati osservati si ottengono eliminado un'osservazione per volta.
Siano *n* e *m* le ampiezze campionarie delle due fonti dei dati $x=(x_1,x_2,\ldots,x_n)$ e $y=(y_1,y_2,\ldots,y_m)$e sia $\hat{\theta} = s(x,y)$ la stima del parametro d'interesse. Complessivamente è possibile ottenere *n* campioni jack-knife di ampiezza *n*-1 dalla prima fonte dei dati e *m* campioni jack-knife di ampiezza *m*-1 dalla seconda fonte dei dati.
Si definiscono di seguito l'i-esimo campione jack-knife della prima fonte dei dati e il k-esimo campine jack-knife della seconda fonte dei dati :

$$x_{(i)} = (x_1,\ldots, x_{i-1}, x_{i+1},\ldots x_n)$$

e

$$y_{(i)} = (y_1,\ldots, y_{k-1}, y_{k+1}\ldots, y_m).$$

Attraverso i campioni jack-knife (one-deleted) è possibile stimare lo standard error jack-knife, il coefficiente di variazione jack-knife, la distorsione jack-knife al fine di valutare l'affidabilità della stima del parametro.

Le seguenti espressioni generalizzano a due fonti, i concetti espressi in 'An introduction to the bootstrap' di Efron e Tibshirani(1993).
La distorsione jack-knife si stima attraverso la differenza del valore atteso delle repliche jack-knife $\hat{\theta}(.)$ e della stima $\hat{\theta}$

$$\hat{bias}_{jack} = (n-1)(m-1)(\hat{\theta}(.) - \hat{\theta}) \qquad (7.46)$$

dove

$$\hat{\theta}(.) = \sum_{i=1}^{n}\sum_{k=1}^{m_y}\frac{\hat{\theta}_{(i,k)}}{nm}, \qquad (7.47)$$

è il valor medio delle repliche jack-knife ('one-deleted').
E' immediata la generalizzazione al corrispondente standard error e coefficiente di variazione jack-knife

$$\hat{SE}_{jack} = \left[\frac{(n-1)(m-1)}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(\hat{\theta}_{(i,k)} - \hat{\theta}(.)\right)^2\right]^{\frac{1}{2}} \qquad (7.48)$$

$$\hat{CV}_{jack} = \frac{\hat{SE}_{jack}}{\hat{\theta}(.)}. \qquad (7.49)$$

Si può così ottenere uno stimatore corretto del parametro d'interesse:

$$\begin{aligned} &\overline{\theta} = \hat{\theta} - \hat{bias}(\theta) \\ &\hat{bias}(\theta) = \hat{\theta}^* - \hat{\theta} \qquad (7.50) \\ &\overline{\theta} = 2\hat{\theta} - \hat{\theta}^*. \end{aligned}$$

Il confronto tra standard error bootstrap e standard error jackknife può essere utile per capire quanto 'smooth' sia una statistica: chiaramente il ricampionamento jack-knife è meno efficiente del ricampionamento bootstrap perchè adopera solo un'informazione limitata della statistica d'interesse. In pratica se $\hat{SE}_{jack} \approx \hat{SE}_{bootstrap}$ allora l'IC studentizzato per θ è accurato ed affidabile: non è quindi necessario utilizzare delle trasformazioni del parametro che rendono costante la varianza di θ.

8 Risultati: Stime puntuali dei coefficienti di trasmissione, stime intevallari “profilo”, modelli “migliori”

8.1 Introduzione

In questo capitolo vengono innanzi tutto presentati i principali risultati riguardanti le stime puntuali dei coefficienti di trasmissione per una serie di modelli basati su alcune matrici “fondamentali”, si sviluppano procedure inferenziali mediante la tecnica degli intervalli di confidenza profilo, e infine si confrontano le prestazioni dei vari modelli considerati nella spiegazione dei dati di infezione mediante le misure di bontà di adattamento (AIC,BIC) descritte nel capitolo precedente.
Successivamente effettuiamo una discussione basata sul concetto relativamente recente di “inferenza multi-modello”, utile in presenza di elevata indecisione nella scelta di un modello tra vari possibili candidati.
In particolare si considerano i modelli riportati nella Tabella 8.1.Questi modelli si possono raggruppare in due tipi di categorie: (a) la classe dei modelli del tipo da noi definito “ad informazione esaustiva” (tipo “M”), (b) la classe dei modelli del tipo “ad informazione non esaustiva” (tipo “C”). Per modelli a “informazione esaustiva” intendiamo modelli basati su stratificazioni dei contatti che costituiscano delle partizioni esaustive dei dati di contatto. Oltre al modello di base (“M1”), con un singolo coefficiente di trasmissione per tutti i tipi di contatti abbiamo considerato una serie di modelli “multi-q”. Per esempio il modello “M2” è un modello a due coefficienti di trasmissione basato sulla stratificazione dei contatti in contatti di tipo “intimo” e contatti di tipo “non intimo”. Questo implica assumere che entrambe le tipologie di contatti siano rilevanti per la trasmissione dell’infezione, sebbene ci aspettiamo che i valori dei rispettivi coefficienti possano risultare molto differenti come grandezza, e che eventualmente qualcuno di essi possa risultare statisticamente non significativo. Similmente abbiamo considerato i modelli con tre parametri “M3” (contatti stratificati in base alla durata), “M4” (contatti stratificati in base al luogo in cui il contatto è stato registrato), “M5” (contatti stratificati in base alla frequenza degli episodi di contatto).
Abbiamo inoltre considerato una serie di modelli one-q ad informazione (o “stratificazione”) non esaustiva. Per modelli ad “informazione non esaustiva” intendiamo modelli basati su stratificazioni dei contatti che costituiscano delle partizioni non esaustive dei dati di contatto. Abbiamo definito così dei modelli in cui solo una parte dei contatti campionati viene effettivamente utilizzata per creare la matrice del numero medio di contatti specifica per il modello. Una stratificazione non esaustiva equivale a considerare il problemi di stima con vincoli di tipo zero su alcuni coefficienti di trasmissione, implicanti la ipotesi a priori che una certa tipologia di contatto sia irrilevante per la trasmissione. Un problema ulteriore è che stratificazioni non esaustive molto “nidificate” hanno lo svantaggio di generare matrici di contatto progressivamente sempre più stocastiche. Pertanto le procedure inferenziali bootstrap del capitolo successivo si limiteranno a considerare i problemi fondamentali basati sulle stratificazioni esaustive.

Tabella 8.1. I vari modelli analizzati e i corrispondenti coefficienti di trasmissione,

Modello	*Descrizione*	*Tipo di stratificazione contatti*	*Numero di coefficienti di trasmissione*
M1 (Contatti complessivi)	***Tutti i tipi di contatti (q_1)***	***Esaustiva***	***1***
M2 (Contatti per intimità)	***"Intimi" (q_1)*** ***"Non intimi" (q_2)***	***Esaustiva***	***2***
M3(Contatti per durata)	***Contatti per durata: D< 15 minuti (q_1), 15 <D<60 minuti (q_2)*** ***D> 60 minuti(q_3)***	***Esaustiva***	***3***
M4(Contatti per luogo)	***Casa(q_1)*** ***Lavoro(q2),*** ***Scuola(q3),*** ***Altro(q4)***	***Esaustiva***	***4***
M5(Contatti per frequenza)	***Giornalieri(q1),*** ***Settimanali(q2),*** ***Occasionali(q3)***	***Esaustiva***	***3***
C1(Contatti "solo fisici")	***"Intimi" (q_1)***	***Non Esaustiva***	***1***
C2:Contatti fisici > 15min .(q1)	***Prossimità(fisica)*** ***Durata(>15 minuti)*** ***(q1)***	***Non Esaustiva***	***1***
C3(Contatti fisici e non fisici > 1h.)	***Prossimità(fisica e non fisica)*** ***Durata(>1h)*** ***(q1)***	***Non Esaustiva***	***1***
C4(Contatti fisici > 15 min. e contatti non fisici > 1 h.(q1))	***Prossimità(fisica)*** ***Durata(>15 min)*** ***+*** ***Prossimità(non fisica)*** ***Durata(>15 min)*** ***(q1)***	***Non Esaustiva***	***1***

.

8.2 Risultati

8.2.1 Risultati Inferenza Modelli

I principali risultati sono riportati nella Tabella 8.2 e nella Tabella 8.3: si può osservare che il confronto dell'AIC e del BIC tra i modelli indica come "migliore" il modello M4 (contatti per luogo) anche se la bontà dell'adattamento è simile per i modelli per prossimità dei contatti M2 (fisici/non fisici) e quello per frequenza del contatto M5. Questo conferma congetture tradizionali per cui le infezioni dei bambini si trasmettono soprattutto a scuola ed in famiglia. Non sembrano rilevanti i contatti non fisici, a lavoro ,di breve durata e poco frequenti.

Posso essere rilevanti per la trasmissione della varicella anche i contatti giornalieri e con durata maggiore a 15 minuti, i contatti nel tempo libero. L'asimmetria di molti degli intervalli di confidenza calcolati per i coefficienti di trasmissione è sempre compresa tra il 2,5 e il 5 %, ovvero è abbastanza ragionevole assumere che asintoticamente la distribuzione dei coefficienti di trasmissione sia Gaussiana, fatta eccezione per i casi in cui il vincolo di non-negatività del

coefficiente rende particolarmente asimmetrico l'IC profilo risultante (q prossimi al valore critico 0). **L' R_0 stimato per ognuno dei modelli considerati si ottiene combinando linearmente gli autovalori dominanti di ognuna delle matrici di contatto usate nel modello**: la Normalità dell'IC profilo dei coefficienti di trasmissione viene ribadita nell'IC per il tasso di riproduzione di base R_0. L'adattamento delle prevalenze attese di immuni in equilibrio ai dati di sieroprevalenza osservati è molto simile tra i modelli, in Fig.8.1 e Fig.8.2 in sono riportati rispettivamente gli adattamenti ai dati di sierologia per i modelli ad informazione esaustiva e non esaustiva considerati.

Tabella 8.2: Stime puntuali IC prifilo dei coefficienti di trasmissione e di R_0 ;criteri di 'model selection'

Modello		IC profile q 95%	q	IC profile R0 95%	R0	Deviance	AIC	BIC
M1	q1	**(0,0308;0,0332)**	**0,032**	**(4.54 ; 4.90)**	**4,72**	**176,49**	**2,03E+03**	**2,03E+03**
M2	q1	**(0,0418;0,0449)**	**0,0433**	**(3,448;3,71)**	**3,57**	**144,53**	**2000,3**	**2011,9**
Proximity	q2	**(0;0,0029)**	**0,0001**					
M3	q1	**(0;0,0094)**	**0,0000**	**(3,864;4,13)**	**3,9916**	**158,88**	**2,02E+03**	**2,03E+03**
Duration	q2	**(0,0221;0,0449)**	**0,0330**					
	q3	**(0,0334;0,0362)**	**0,0348**					
M4	q1	**(0,0053;0,0215)**	**0,0129**	**(3,52;3,67)**	**3,58**	**125,31**	**1,99E+03**	**2,01E+03**
Location	q2	**(0;0,0496)**	**0,0000**					
	q3	**(0,0404;0,0437)**	**0,0420**					
	q4	**(0;0,0173)**	**0,0068**					
M5	q1	**(0,0356;0,0384)**	**0,0369**	**(3,93;4,24)**	**4,08**	**135,71**	**1,99E+03**	**2,01E+03**
Frequency	q2	**(0;0,0092)**	**0,001**					
	q3	**(0;0,0073)**	**0**					

Tabella 8.3:IC profilo per i coefficienti di trasmissione q e criteri di 'model selection'

Modello		IC profile q 95%	q	IC profile R0 95%	R0	Deviance	AIC	BIC
C1	q1	**(0,1083;0,1228)**	**0,1152**	**(8,94;10,13)**	**9,5109**	**367,0692**	**2,22E+03**	**2,23E+03**
Phisical								
C2	q1	**(0,1145;0,1228)**	**0,1296**	**(8,14;9,21)**	**8,6509**	**382,03**	**2,23E+03**	**2,24E+03**
Phisical>15min								
C3	q1	**(0,0956;0,1078)**	**0,1014**	**(8,77;9,89)**	**9,3038**	**444,02**	**2,30E+03**	**2,30E+03**
Proximity>1hour								
C4	q1	**(0,0881;0,0995)**	**0,0936**	**(9,07;10,24)**	**9,63**	**419,44**	**2,27E+03**	**2,28E+03**
Phisical>15min; Not Phisical>1hour								

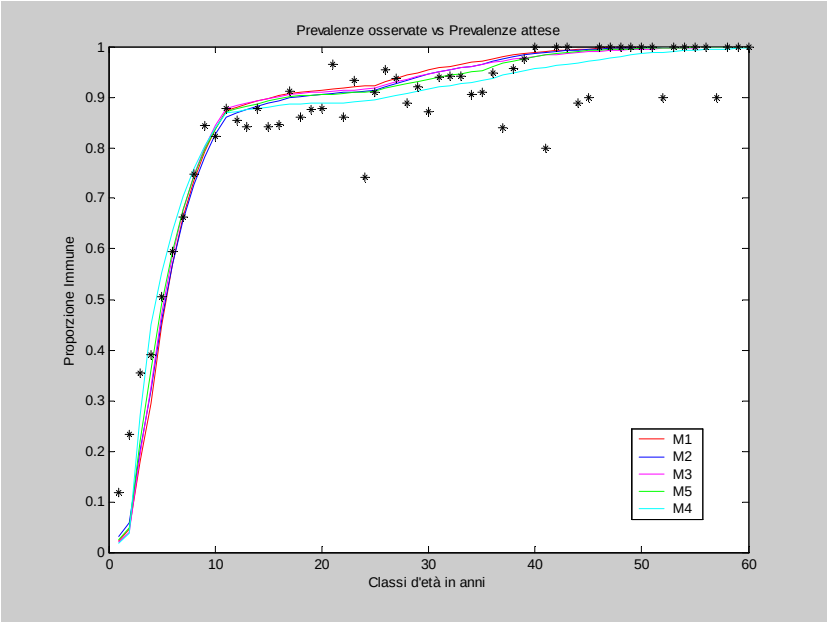


Fig. 8.1.Italia. Adattamento dei modelli ad informazione esaustiva ai profili di sieroprevalenza osservati

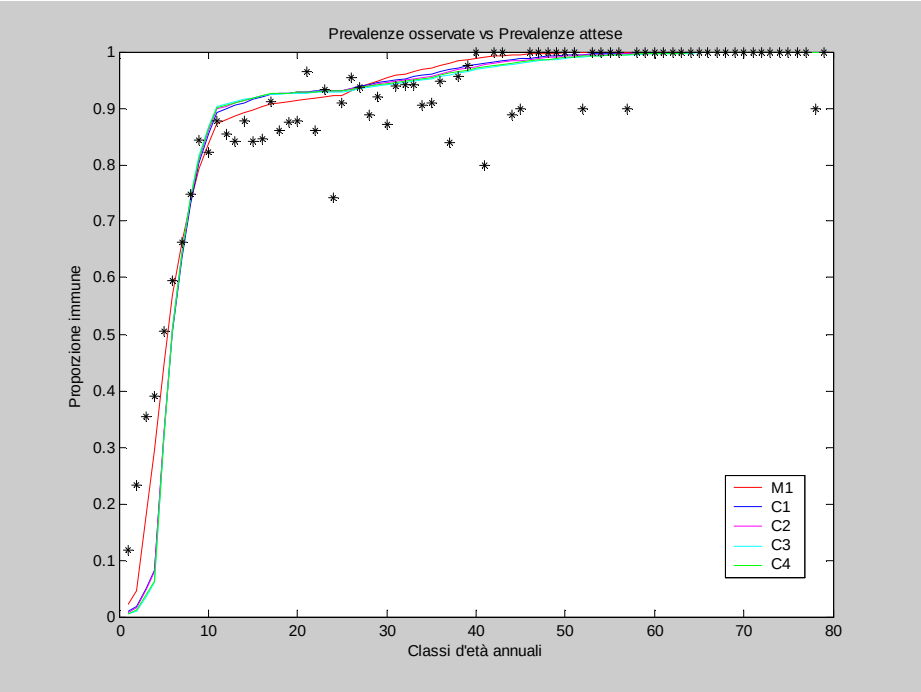


Fig. 8.2.Italia. Modelli one-q (M1+modelli ad informazione non esaustiva):Adattamento delle prevalenze attese alle prevalenze osservate .

Per i modelli one-q e per il modello M1(contatti complessivi)la log-verosimiglianza è definita sul piano. Per i modelli ad informazione non esaustiva analizzati la log-verosimiglianza, raggiunto il valore massimo per poi decrescere molto lentamente rispetto alla verosimiglianza del modello M1 .

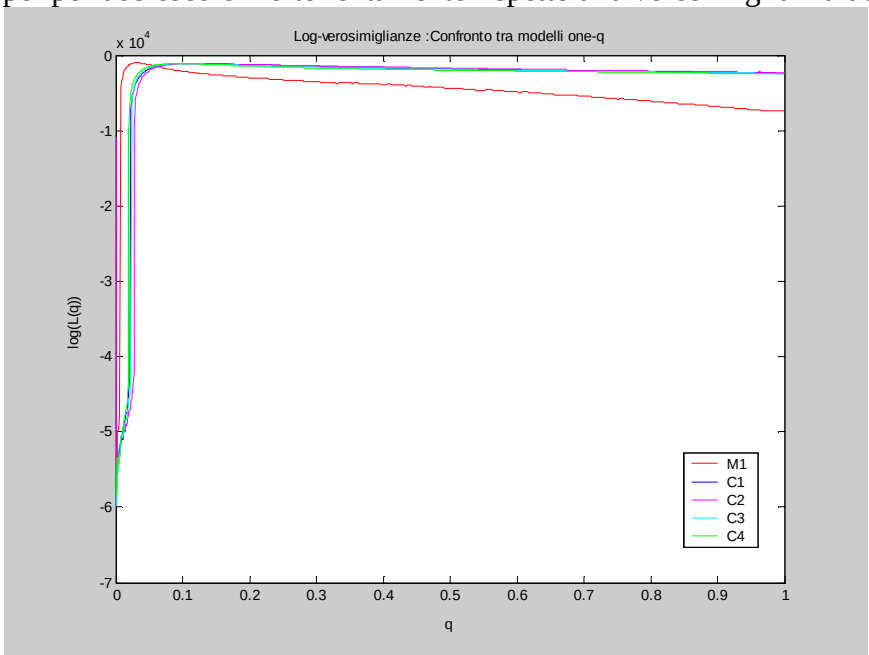


Fig. 8.3.Italia.Log-verosimiglianze dei coefficienti di trasmissione per i Modelli one-q

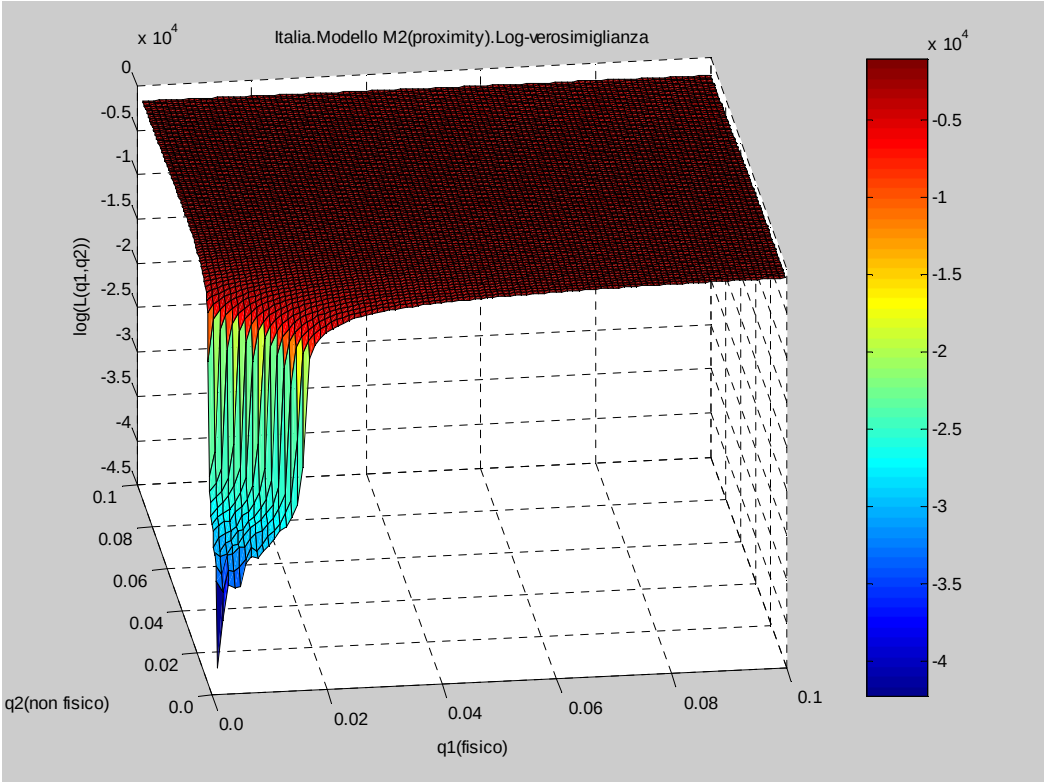


Fig. 8.4.Italia.Modello M2(contatti per prossimità) Log-verosimiglianza

8.3 Inferenza multimodel

Per i modelli considerati in precedenza sono stati calcolati i pesi e le differenze di Akaike .

Tabella 8.4:Inferenza multi-modello

Modello	Δ(Diff. Akaike)	W (Pesi Akaike)
M1	45,5	0,0250
M2	15,2	0,1200
M3	31,6	0,0550
M4	0	0,6695
M5	9,9	0,1305
C1	235,9	0,0000
C2	242,8	0,0000
C3	311,9	0,0000
C4	287,9	0,0000

L'evidenza è a favore del modello M4: i contatti per luogo sono i più importanti nella trasmissione della varicella;c'è una piccola evidenza a favore dei contatti per prossimità e di quelli per frequenza. Chiaramente la forte evidenza a favore del modello per luogo dei contatti rende poco utile l'inferenza multimodello, che invece ha la sua utilità in situazioni di incertezza nella scelta di un modello migliore tra i candidati.

E' inoltre possibile calcolare la matrice dei rapporti dell'evidenza

(w_i /w_j): gli elementi sulla diagonale(pari a 1) non sono informativi, per tutti gli altri valori quanto più è grande l' 'Evidence Ratio ' tanto più è forte l'evidenza del modello i sul modello j .

Il modello M4 (contatti per luogo) è il più plausibile; il modello M5(contatti per frequenza) ha un'evidenza superiore al modello M2(contatti per prossimità).

9 Risultati:Inferenza Bootstrap sui parametri di trasmissione

9.1 Introduzione

In questo capitolo utilizziamo il bootstrap non parametrico per ricampionare le due fonti d'incertezza del modello – quella connessa con i dati di contatto e quella connessa con il dato sierologico – al fine di caratterizzare la variabilità delle stime dei coefficienti di trasmissione q. Il nostro scopo è quello di riuscire a caratterizzare l'incertezza intrinseca di ciascuna delle due fonti considerate separatamente, e di come queste incertezze si combinano quando le consideriamo congiuntamente. A tale scopo considereremo tre distinti livelli:

1. ricampionamento dei dati di sierologia condizionatamente alla matrice dei contatti sociali (ovvero assumendo assenza di incertezza nel dato di contatto), al fine di valutare l'incertezza intrinseca del dato di infezione.
2. ricampionamento dei dati dei contatti sociali Polymod, condizionatamente ai profili di seroprevalenza osservati nell'indagine Esen2, al fine di valutare l'incertezza intrinseca del dato di contatto.
3. ricampionamento congiunto di entrambe le fonti dei dati, al fine di valutare le modalità con cui le due fonti si combinano quando considerate congiuntamente.

Per ciascuna di queste fasi l'aspetto centrale del lavoro sarà costituito dalla determinazione del numero minimo di repliche bootstrap ("isolate" per i livelli A,B, oppure congiunte per il livello C) necessarie affinché lo standard error bootstrap dei coefficienti di trasmissione sia approssimabile a quello ideale. Successivamente forniremo una analisi comparativa dei risultati relativi agli standard error ideali e valutazioni sulla velocità di convergenza al valore ideale. Infine forniremo gli intervalli di confidenza bootstrap per i coefficienti di trasmissione, per i differenti approcci discussi nel capitolo 7.

Come già anticipato nel capitolo precedente le analisi in oggetto verranno sviluppate per i modelli del tipo "M": M1 (contatti totali), M2 (contatti per prossimità), M3(contatti per durata), M4 (contatti per luogo), M5 (contatti per frequenza).

In generale ,come ora documenteremo ,le analisi svolte documentano i seguenti fatti principali:
A) Le repliche bootstrap mostrano che le stime dei coefficienti di trasmissione meno rilevanti (es: contatti non fisici, contatti minori di 15 minuti,contatti a scuola, contatti sporadici/occasionali) presentano spesso una maggiore variabilità rispetto a quelle dei coefficienti più rilevanti. Si considerano rilevanti i coefficienti di trasmissione(significativamente diversi da zero) .Gli intervalli di confidenza costruiti su questi parametri risulteranno in definitiva poco affidabili .
B) Alcune tipologie di contatto non riescono a spiegare i dati di sierologia e quindi le repliche bootstrap dei coefficienti di trasmissioni associate a tale tipologia hanno

- ✓ standard error bootstrap che convergono per numeri di repliche maggiori rispetto a quelle necessarie per i parametri più rilevanti ,
- ✓ distorsione bootstrap a volte non trascurabili,
- ✓ coefficienti di variazione bootstrap elevati .

C)Al contrario i coefficienti di trasmissione più rilevanti sono caratterizzati dal fatto che la convergenza allo standard error ideale si verifica per un numero minore di repliche , danno luogo a standard error bootstrap dei coefficienti di trasmissione più affidabili (coefficienti di variazione bootstrap piccoli e distorsione bootstrap trascurabile).

Gli intervalli di confidenza costruiti su coefficienti di trasmissione più rilevanti risultano quindi maggiormente affidabili. Fatta eccezione per il modello M5(contatti per frequenza), i dati sui contatti sociali hanno sempre il peso maggiore sulla variabilità complessiva del modello considerato rispetto ai dati di seroprevalenza : gli standard error ideali dei coefficienti di trasmissione condizionati ai profili di seroprevalenza osservati sono sempre più grandi di quelli condizionati alle matrici di contatto osservate e quasi sempre superiori di 1 o 2 ordini di grandezza .

9.2 Numero di repliche necessarie alla convergenza agli standard error ideali

Di seguito viene presentata la Tabella 9.1 in cui, per i modelli analizzati (prima colonna), si valutano i numeri di repliche necessari alla convergenza dello standard error ideale dei parametri di trasmissione nel caso si consideri una sola fonte d'incertezza (seconda e terza colonna) e nel caso in cui le fonti d'incertezza considerate siano due (quarta colonna).Si indicheranno con B il numero di ricampionamenti dei dati di contatto e con P il numero di ricampionamenti dei dati di sierologia per la restante parte del capitolo.
La convergenza viene valutata graficamente: al variare del numero di ricampionamenti bootstrap (nel caso di fonte unica o di fonte multipla d'incertezza) si valuta l'andamento dello standard error bootstrap e si stabilisce il numero di repliche necessarie a garantire che ,per quel numero di repliche,lo standard error bootstrap sia una buona approssimazione dello standard error ideale per il corrispondente coefficiente di trasmissione analizzato.

Tabella 9.1:Numero di repliche necessarie alla convergenza al valore ideale dello standard error bootstrap dei coefficienti di trasmissione q

Modello	Convergenza Fonte incertezza: Sierologia (P)	Convergenza Fonte incertezza: Contatti (B)	Convergenza Fonte incertezza: Sierologia e contatti (P,B)
M1	800	1000	(50,300)
M2	800	1250	(100,150)
M3	800	800	(100,300)
M4	600	600	(50,200)
M5	600	800	(200,200)

LA Tabella 9.1 suggerisce che quando si considera un'unica fonte di incertezza, al fine di conseguire la convergenza degli standard error bootstrap dei coefficienti di trasmissione q al loro valore ideale è quasi sempre sufficiente un numero di repliche in un range da 600 a 1000 con la sola eccezione del modello dei contatti intimi. Questo risultano non sembra dipendere dal numero di parametri da stimare. Infine l'ultima colonna della tabella mostra il numero di repliche P e B minimo necessarie ad ottenere la convergenza degli standard error bootstrap allo standard error ideale: è quindi possibile che alcuni coefficienti di trasmissione convergano anche per un numero minore di repliche del corrispondente standard error bootstrap. A supporto della Tabella 9.1, vengono presentati , per il modello M1,alcuni grafici(Fig. 9.1,Fig. 9.2,Fig. 9.3) che mostrano la convergenza dello standard error bootstrap allo standard error ideale per i parametri di trasmissione.

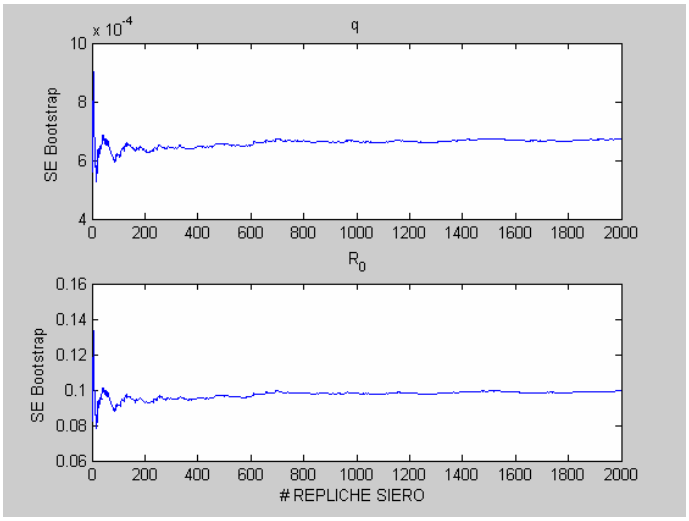


Fig. 9.1.Modello M1;Ricerca dello standard error bootstrap ideale per q ed R0 al variare del numero di repliche dei dati sierologici

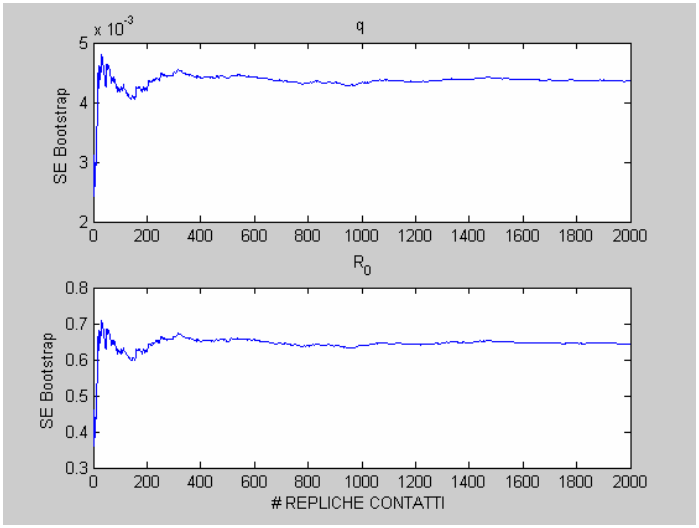


Fig. 9.2.Modello M1; Standard error bootstrap dei parametri al variare del numero di repliche dei dati di contatto

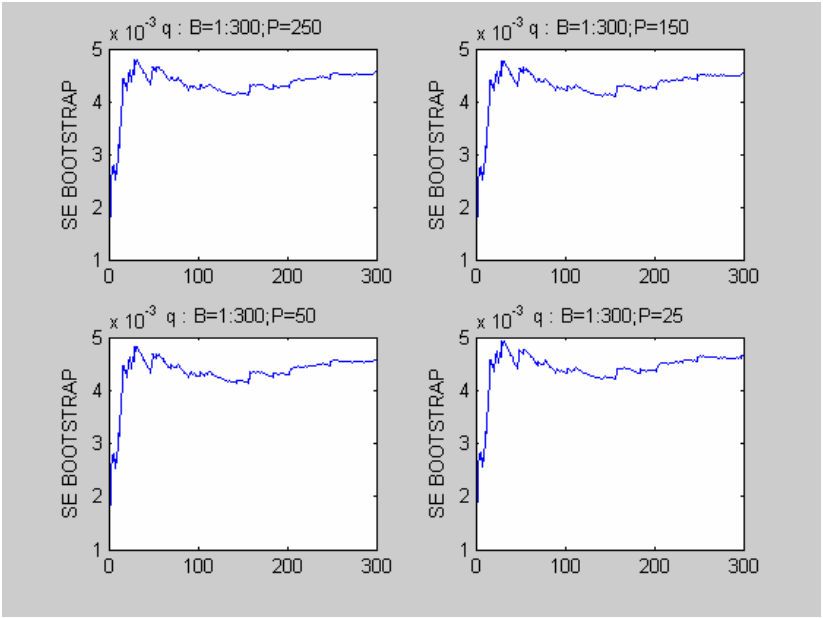


Fig. 9.3Modello M1;Standard error bootstrap del coefficiente di trasmissione q al variare di B (numero di repliche dei dati di contatto),fissato P(numero di repliche dati sierologici)

Entrando più in dettaglio sui fattori che influenzano la velocità di convergenza risultano i seguenti fatti: in primo luogo quando analizziamo la velocità di convergenza per ogni singolo parametro, questa sembra aumentare all'aumentare della sua rilevanza .In Fig. 9.4 è analizzato l'andamento dello standard error bootstrap quando condizionatamente ai dati sierologici osservati si ricampionano i dati sui contatti sociali per il modello M2 (contatti per prossimità) :si nota che la maggiore variabilità del parametro q2 (contatti non fisici) rispetto al parametro q1(contatti fisici), riscontrabile analizzando i rispettivi coefficienti di variazione dei 2 parametri , inevitabilmente si traduce in un numero maggiore di repliche necessarie alla convergenza dello standard error bootstrap del parametro q2 rispetto alle repliche necessarie per q1.

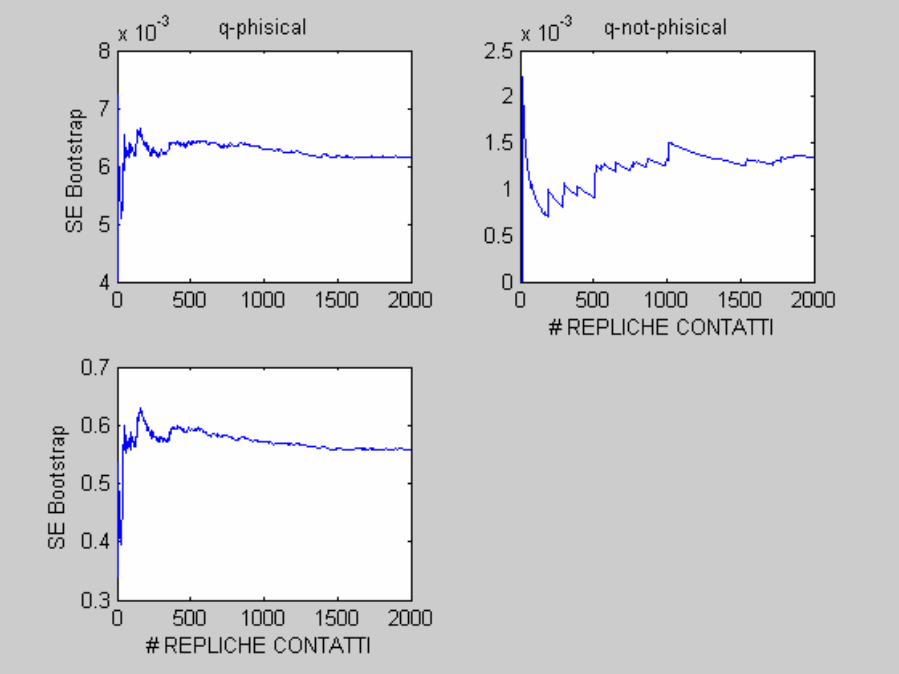


Fig. 9.4.Modello M2;Standard error bootstrap dei parametri di trasmissione(q1,q2,R0) al variare del numero B di repliche dei dati di contatto

In secondo luogo ,quando consideriamo un assegnato modello e ricampioniamo separatamente rispetto alle due fonti d'incertezza allora la presenza di una forte differenza tra se ideali condizionati ad una delle due fonti di incertezza piuttosto che all'altra si traduce in una maggiore velocità di convergenza ; per contro, quando le differenze tra le componenti di variabilità si fanno meno evidenti , la velocità di convergenza è minore.
Se si considera il modello dei contatti complessivi M1 lo standard error ideale condizionato alla matrice dei contatti (la fonte d'incertezza è il dato sierologico) dell'unico coefficiente di trasmissione (q) è dell'ordine di grandezza di 10^{-4}(Tabella 9.2,colonna2),mentre lo standard error ideale condizionato ai dati di sierologia(la fonte di incertezza è il dato sui contatti sociali) dello stesso parametro è dell'ordine di grandezza di 10^{-3}(Tabella 9.2,colonna3):la fonte dei contatti sociali è la più rilevante nello spiegare la variabilità complessiva del modello.
Così non avviene nel modello M5 (contatti per frequenza):gli ordini di grandezza degli standard error condizionati prima ad una e poi all'altra fonte dei coefficienti di trasmissione (q1,q2) hanno lo stesso ordine di grandezza (Tabella 9.6,colonna 2 e 3) .Per il modello M5 la variabilità complessiva deriva da entrambe le fonti d'incertezza in misura simile.
Il numero di repliche necessarie alla convergenza dello standard error bootstrap di q1 per il modello M1 sono P*B=15000(P=50,B=300); per il modello M5 il numero di repliche necessarie alla convergenza dello standard error bootstrap dei parametri del modello è P*B=40000 (P=200,B=200).

In particolare gli standard error ideali condizionati prima ad una e poi all'altra fonte, se confrontati , possono spiegare differenti velocità di convergenza degli standard error ideali per le fonti d'incertezza ricampionate congiuntamente.
In questa tesi è stato utilizzato un criterio grafico per ottenere la coppia ottimale (P,B)* nel caso in cui si analizzano le due fonti d'incertezza congiuntamente.
Fissando il numero di repliche per una delle due fonti d'incertezza, si fa variare il numero di repliche bootstrap utilizzate per l'altra fonte,determinando il numero di repliche necessarie alla convergenza .Quindi si itera il procedimento scegliendo un differente numero di repliche dell'altra fonte. Il procedimento viene poi replicato scambiando il ruolo delle due fonti. Questo meccanismo anche se grossolano, sembra molto utile al calcolo del numero di repliche necessario per la convergenza degli se ideali:un criterio numerico può essere facilmente implementato per la risoluzione di questo problema di minimo definendo ad esempio un tolleranza sulla variabilità(es: $|cv_{ideale}-cv_i|<$tolleranza) ed imponendo che, per la distribuzione degli standard error bootstrap derivante dal numero di repliche scelte, il valore dello standard error ideale sia ,ad esempio, il valore modale.

Valutare il numero di ricampionamenti bootstrap necessari ad ottenere gli standard error ideali dei parametri di trasmissione è semplice nel caso in cui la fonte d'incertezza sia unica . Al variare,ad esempio, del numero di ricampionamenti bootstrap dei dati Polymod si calcola il corrispondente valore dello standard error bootstrap : sul piano (se_b,b) con b=1…..B si può facilmente scegliere visivamente o tramite un criterio numerico tale numero .Come si può intuire in Fig. 9.5,bastano poco più di 600 ricampionamenti bootstrap della matrice dei contatti affinché lo standard error bootstrap dell'R_0 per il modello M3 sia approssimabile al valore ideale dello standard error ideale ($B = +\infty$) .
Un criterio numerico può essere facilmente creato, imponendo che una fissata tolleranza sia maggiore del valore assoluto della differenza tra potenziale se ideale e lo standard error bootstrap considerato, e contemporaneamente accertandoci che il potenziale standard error in questione appartenga alla classe modale della distribuzione degli standard error bootstrap .
Questo secondo vincolo può essere sostituito considerando ,ad esempio ,il rapporto tra il numero di volte che lo standard error bootstrap assume il valore ideale e il numero di ricampionamenti bootstrap totali :se tale proporzione supera il ad esempio il 75% allora lo se è ideale.
Si valuta infine il valore b (numero di ricampionamento) per cui tali condizioni sono soddisfatte.

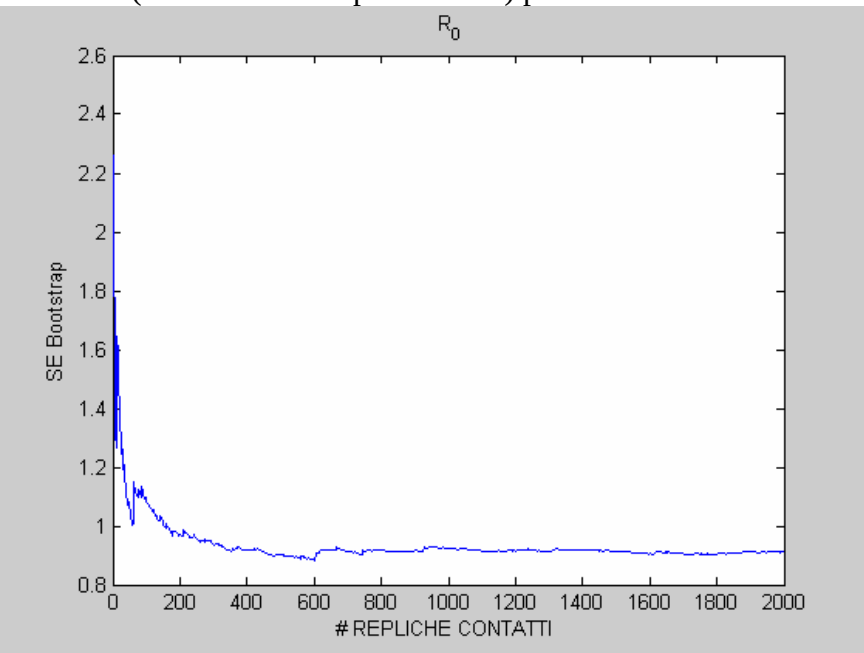


Fig. 9.5.Modello M3:Valutazione del numero di ricampionamenti bootstrap necessari ad ottenere se ideale per R0

Nel caso in cui le fonti di incertezza sono due, il criterio grafico è ancora valido ed utile a valutare il numero di ricampionamenti bootstrap necessari ad ottenere lo standard error ideale .
Si fissa il numero di ricampionamenti bootstrap per una delle due fonti ,mentre si fa variare l'altra fonte dal numero minimo al numero massimo di repliche precedentemente fissato.
In tal modo si valuta il contributo individuale di ognuna delle fonti condizionatamente all'altra in modo da poter stabilire il numero di ricampionamenti minimo della fonte Polymod(B) e della fonte sierologica (P) per ottenere buone approssimazioni dello se ideale dei parametri di trasmissione .

Osservazione:Se in alcuni modelli,ogni parametro a dei valori (B,P) diversi di convergenza ,si utilizza il seguente criterio: per il modello la coppia (B,P)* ottimale sarà il massimo (per B e per P) tra i minimi necessari ad ottenere la convergenza al variare di tutti i suoi parametri.

$$(B,P)^* = (\max(B_{q_i}), \max(P_{q_i}))$$

$$i = 1...n \qquad (9.1)$$

$$n = \#\ parametri$$

Le figure successive mostrano come si giunge al valore ottimo di (B,P)* nel caso di fonte doppia di incertezza .Si valuterà per alcuni modelli la convergenza del tasso di riproduzione di base della varicella .

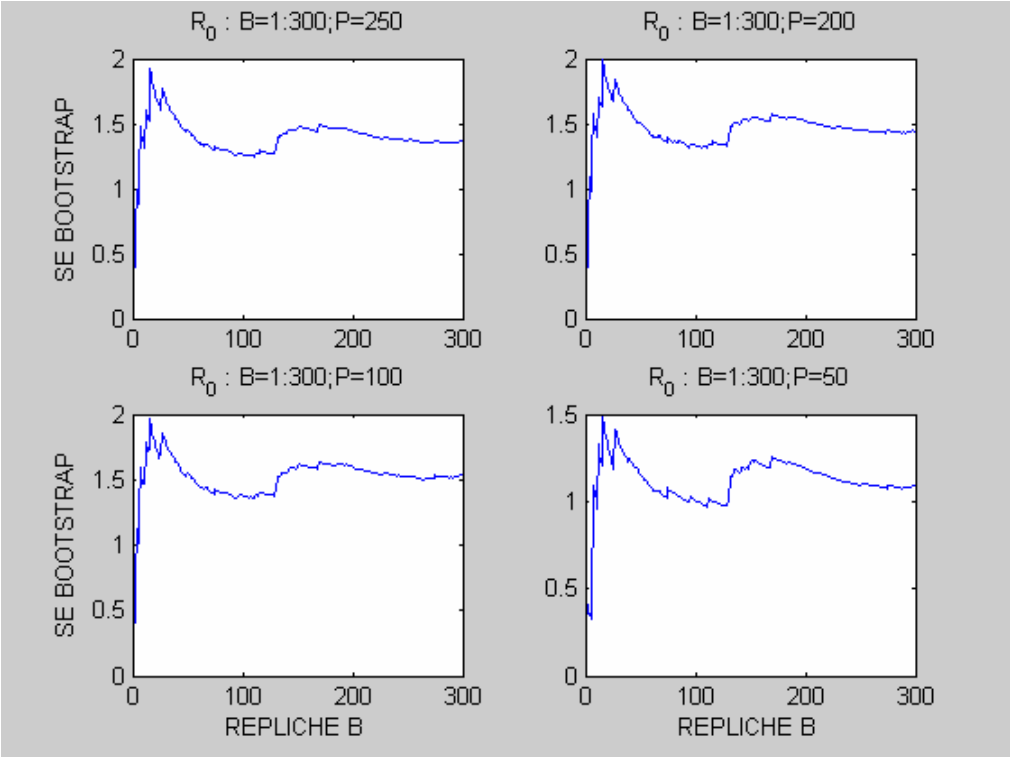


Fig. 9.6.Modello M2;Standard error bootstrap del parametro di trasmissione R0 al variare del numero B di repliche dei dati di contatto fissando P, repliche dei dati sierologici (P=250,200,100,50)

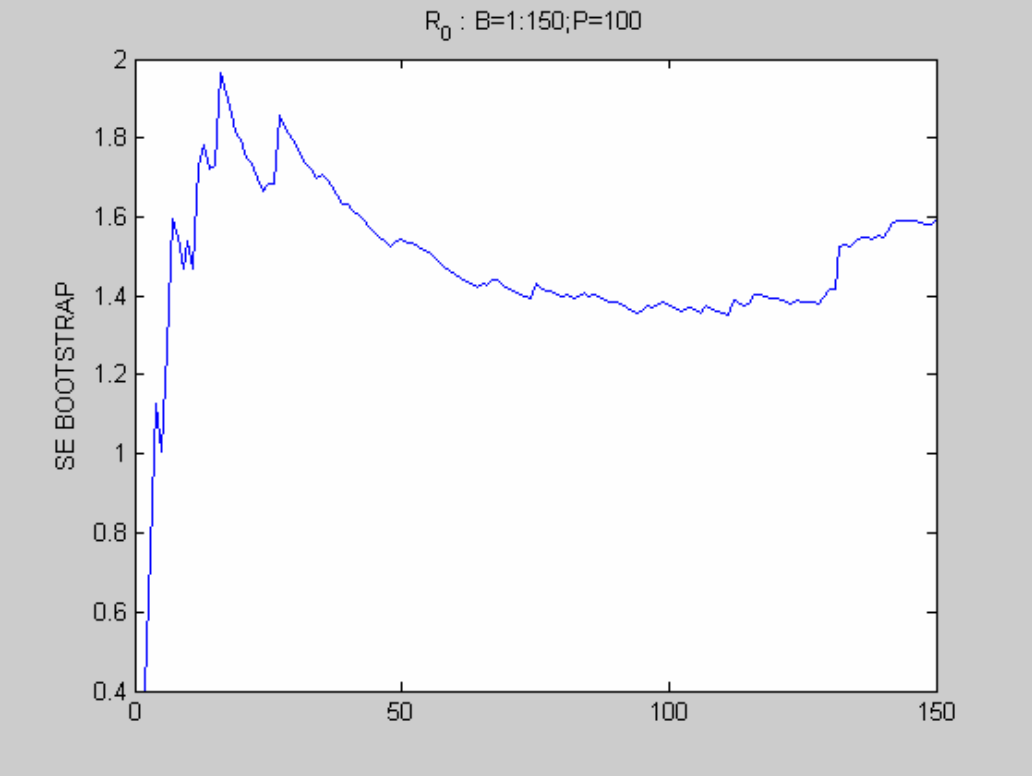


Fig. 9.7: Modello M2;Standard error bootstrap del parametro di trasmissione R0 al variare del numero B di repliche dei dati di contatto fissato P, repliche dei dati sierologici (P=50).(B,P)*=(150,100) è sufficiente a garantire la convergenza allo se ideale di R0

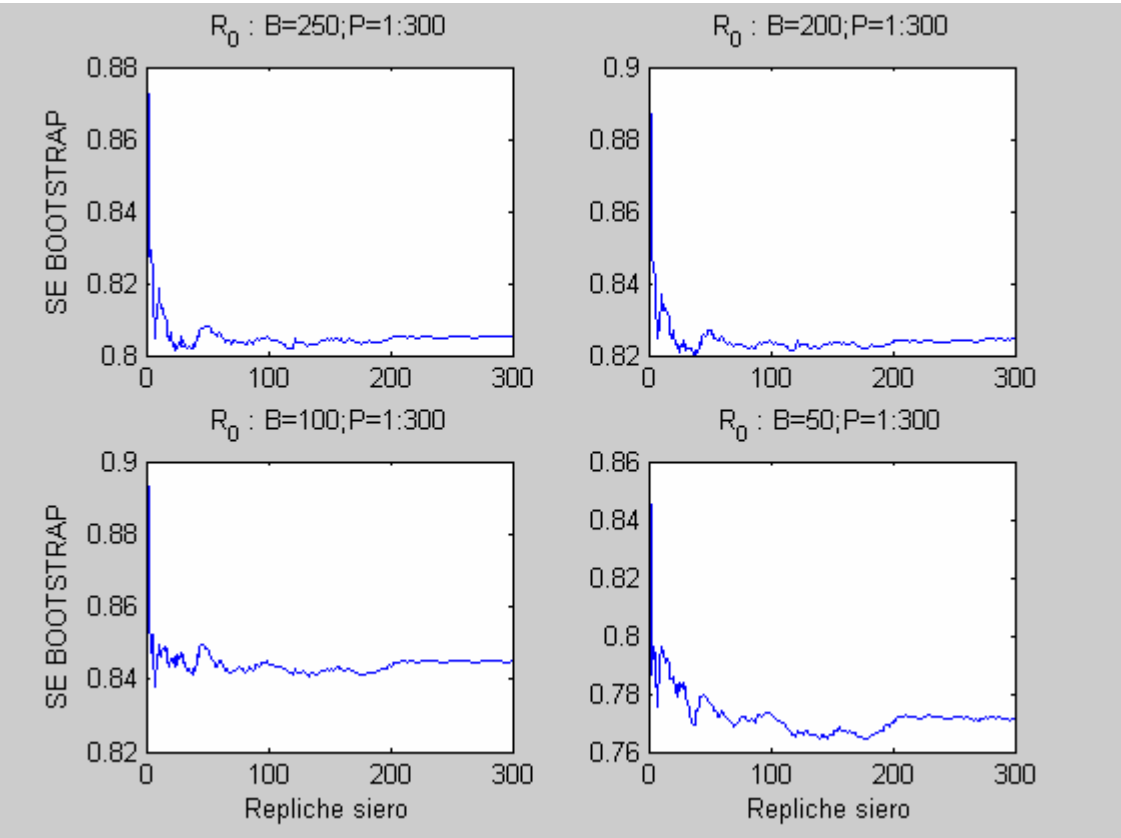


Fig. 9.8.Modello M3:Se bootstrap per R0 al variare del numero P di ricampionamenti dei profili di seroprevalenza ,fissato il numero B=250,200,100,50 di ricampionamenti della matrice dei contatti

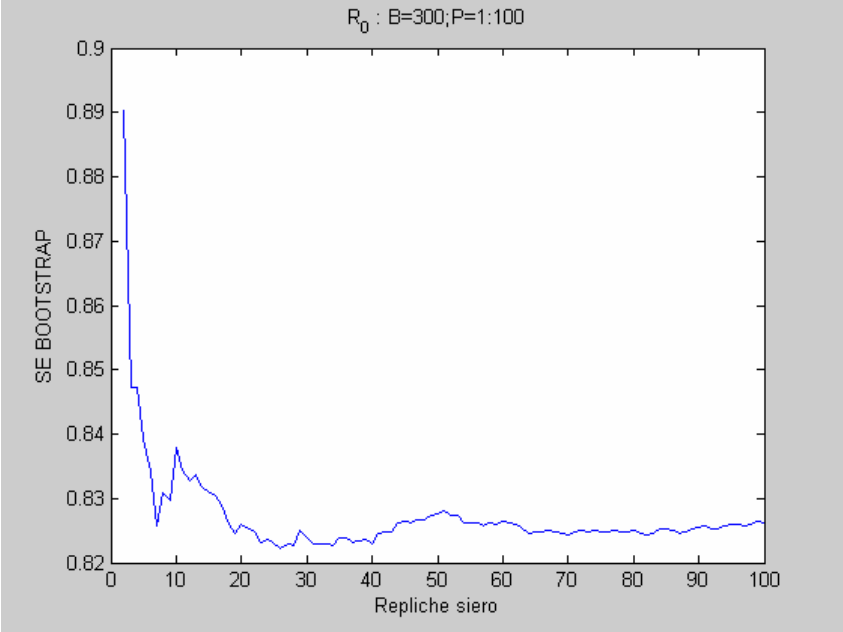


Fig. 9.9.Modello M3:Se bootstrap per B=300 al variare di P=1…100.(B,P)*=(300,100) è sufficiente ad ottenere una buona approssimazione dello se ideale

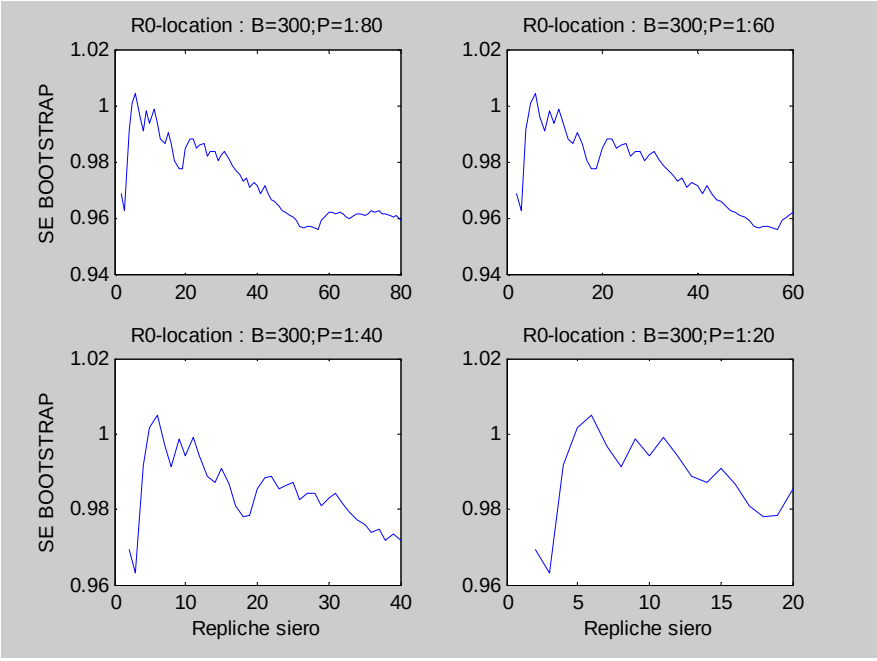


Fig. 9.10.Modello M4:Se bootstrap dell'R0 ,fissato il numero di ricampionamenti dei contatti sociali B=300 ,si fa variare il numero di ricampionamenti dei profili di seroprevalenza.(P*=50)

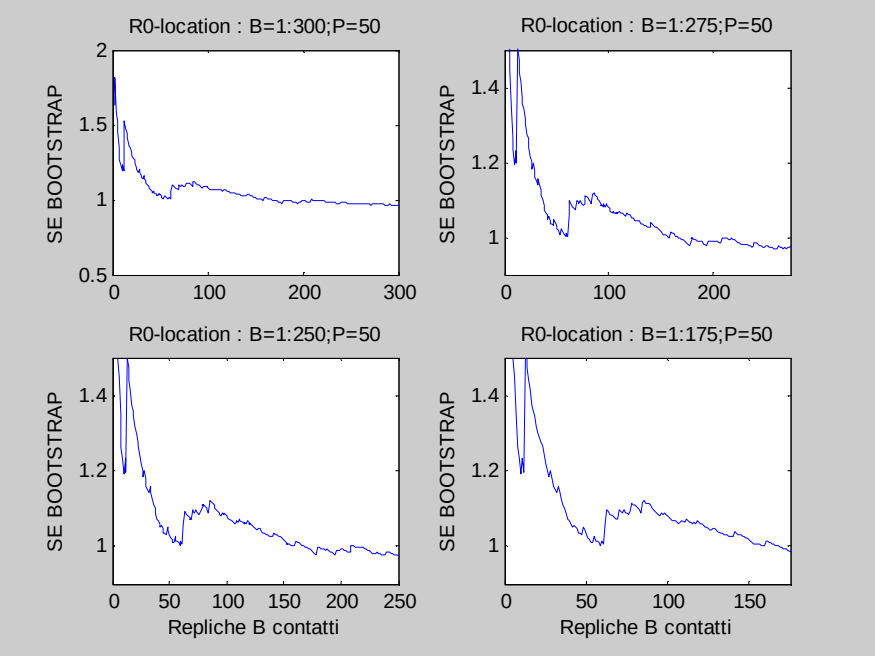


Fig. 9.11. Modello M4:Se bootstrap dell'R0 ,fissato il numero di ricampionamenti dei dati ESEN2 P=50 ,si fa variare il numero di ricampionamenti dei dati Polymod.(B*=200)

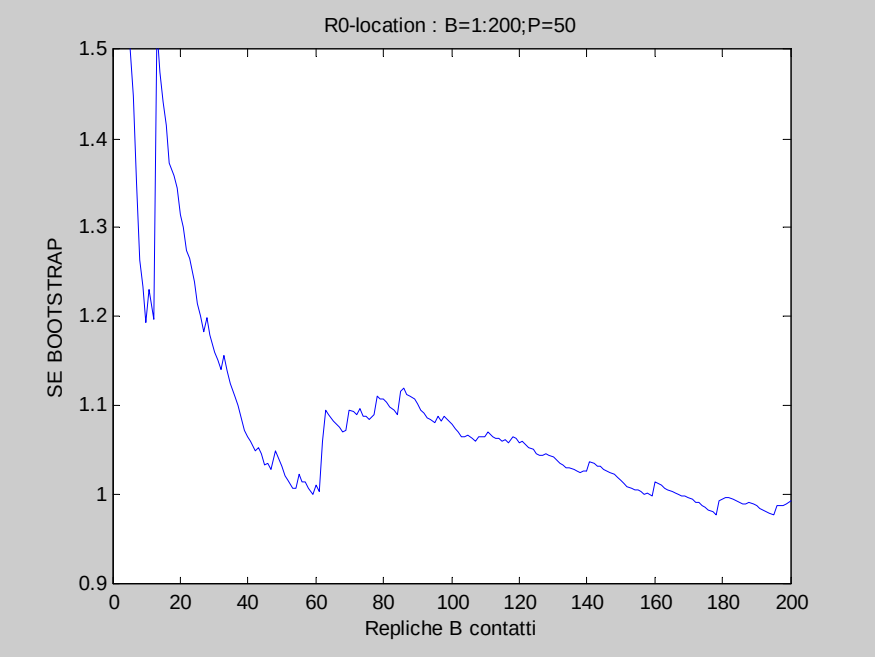
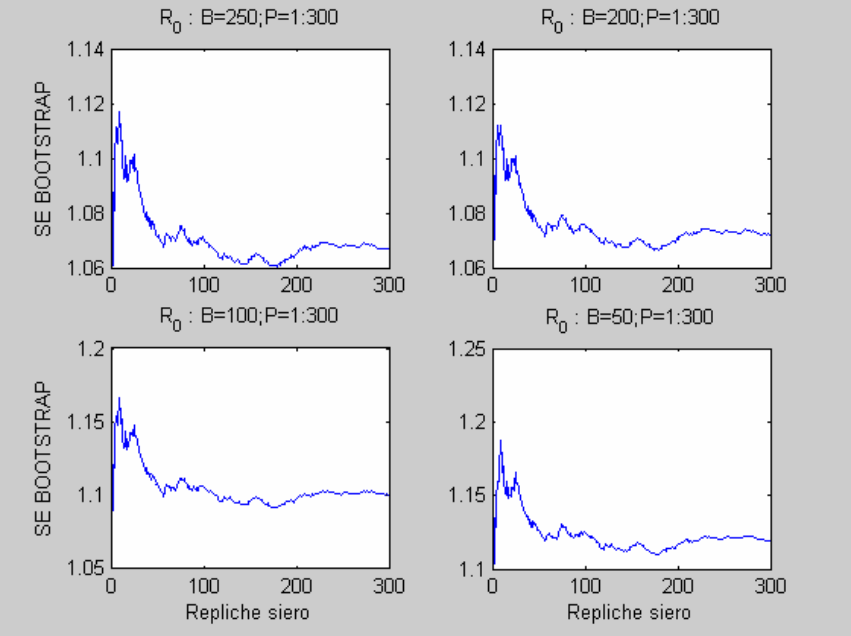


Fig. 9.12 . Modello M4:Se bootstrap dell'R0 ,fissato il numero di ricampionamenti dei dati ESEN2 P=50 ,si fa variare il numero di ricampionamenti dei dati Polymod.(B,P)*=(200,50)



Modello M5:Se bootstrap dell'R0 ,fissato il numero di ricampionamenti dei dati Polymod B=250,200,100,50,si fa variare il numero di ricampionamenti dei profili di seroprevalenza P=1…300 ;(B,P)*=(200,200)

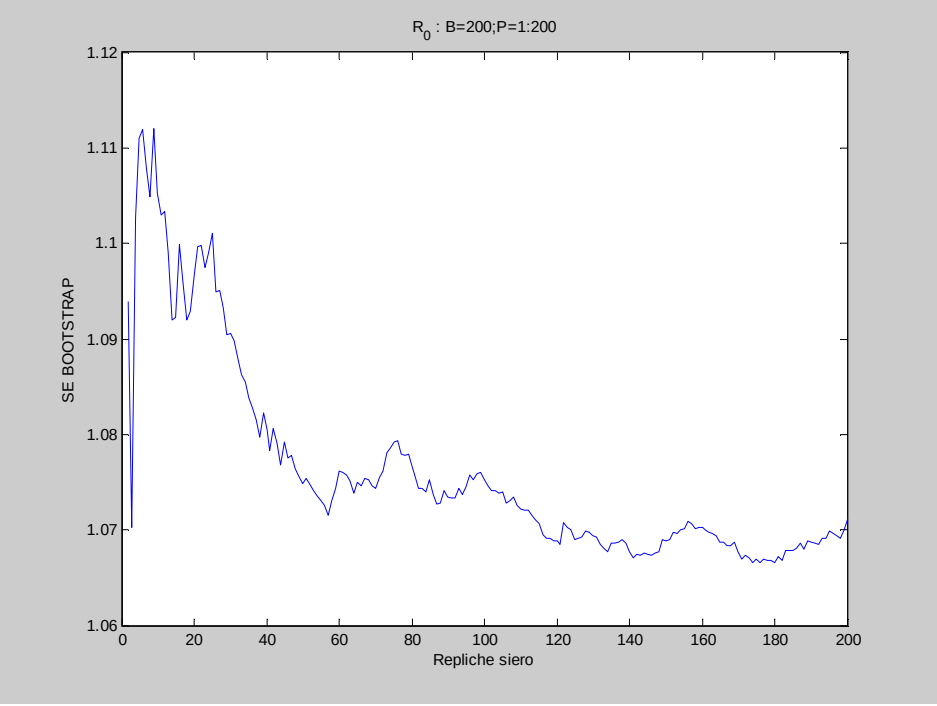


Fig. 9.13.Modello M5:Se bootstrap di R0 ,fissato il numero di ricampionamenti dei dati di contatto B=200,si fa variare il numero di ricampionamenti dei profili di seroprevalenza P=1…200.(B,P)*=(200,200) è sufficiente ad ottenere una buona approssimazione dello se ideale di R0.

I risultati in Tabella 9.1 sono stati ottenuti analizzando gli andamenti dello standard error bootstrap per tutti i parametri dei differenti modelli .
Gli standard error ideali condizionati prima ad una e poi all'altra fonte ,se confrontati , possono spiegare differenti velocità di convergenza degli se ideali per le fonti d'incertezza congiuntamente . Per un modello una forte differenza tra se ideali condizionati ai contatti osservati e se ideali condizionati alle seroprevalenze osservate si traduce in una maggiore velocità di convergenza ; per contro, quando le differenze tra le componenti di variabilità si fanno meno evidenti , la velocità di

convergenza è minore. L'individuazione del numero minimo di repliche bootstrap dei coefficienti di trasmissione mira a fornire indicazioni su come contenere i tempi computazionali mantenendo però il livello desiderato di accuratezza.

9.3 INTERVALLI DI CONFIDENZA BOOTSTRAP PER COEFFICIENTI DI TRASMISSIONE

9.3.1 Premessa metodologica

Riportiamo alcune brevi note sugli accorgimenti tecnici che abbiamo seguito al fine di procedere alla costruzione degli intervalli di confidenza bootstrap garantendo omogeneità e comparabilità dei risultati trovati**.**Gli IC costruiti per i modelli M1-M2-M3-M4-M5 sono stati valutati in presenza delle due fonti d'incertezza :il numero di ricampionamenti scelto per entrambe le due fonti è 300,per un totale di B*P=90000 repliche per ognuno dei coefficienti di trasmissione .
La comparabilità degli IC è quindi garantita dalla constatazione che per quel livello di ricampionamenti tutti i modelli analizzati giungono a risultati di convergenza e dal fatto che tutti gli IC sono fondati su distribuzioni di repliche bootstrap dei parametri di trasmissione di uguale numerosità. L'unica tipologia non confrontabile direttamente alle altre è l'IC studentized che è calcolato per un numero di repliche diverse(B=P=40 ricampionamenti di primo livello e B2=P2=10 ricampionamenti di secondo livello per un totale di 160000 repliche).
Gli intervalli di confidenza bootstrap normali sono costruiti tramite lo standard error ideale dei coefficienti di trasmissione q. Tali intervalli possono molte volte essere migliorati depurandoli dalla loro distorsione bootstrap o attraverso tecniche di calibrazione che , ancora una volta, prevedono l'utilizzo di repliche bootstrap di secondo livello per valutare la copertura reale dell'IC. La debolezza dell'IC nomale bootstrap è l'assunzione di normalità, che spesso non trova riscontro nelle distribuzioni bootstrap delle repliche dei paramentri d'interesse.
In Fig. 9.14 si presentano le distribuzioni delle repliche bootstrap dei parametri di trasmissione per il modello M3(contatti per durata).Si può notare come il coefficiente di trasmissione q1(contatti con durata inferiore a 15 minuti) sia una distribuzione molto asimmetrica: l'assunzione di Normalità non è certo in linea con la distribuzione bootstrap delle repliche.

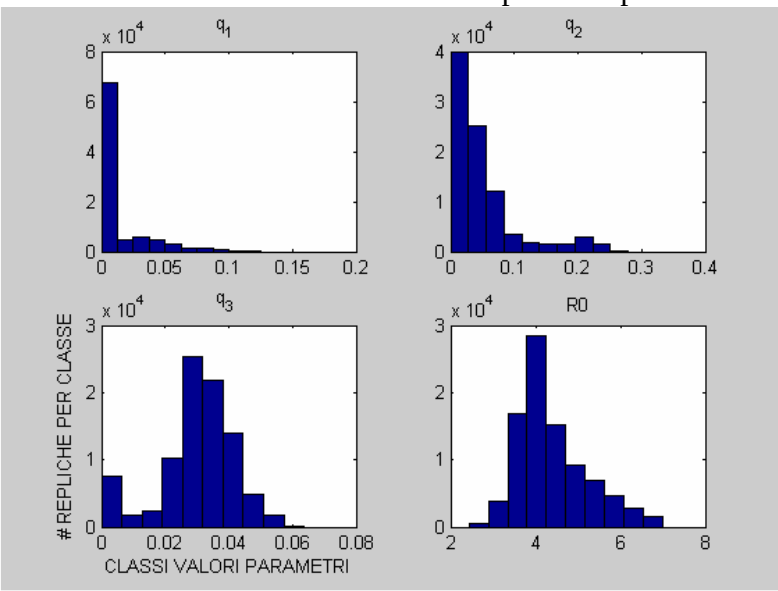


Fig. 9.14.Modello M3(contatti per durata).Distribuzione delle repliche bootstrap dei parametri di trasmissione

L'IC percentile si fonda totalmente sulla distribuzione bootstrap delle repliche dei coefficienti q e non considera l'eventualità di una possibile distorsione nella stima di tali paramentri : come per

l'IC normale , si potrebbe migliorare l'intervallo aggiustandolo per la distorsione bootstrap ,qualora quest'ultima non fosse trascurabile , o si potrebbe migliorarne l'accuratezza attraverso la calibrazione .Le problematiche legate alla costruzione dell'IC bootstrap-t sono ancora una volta dovute ai tempi di calcolo; inoltre questa tipologia d'intervallo risente dei valori erratici delle repliche e quindi si potrebbe migliorare utilizzando delle trasformazioni degli standard error di secondo livello in modo da stabilizzarne la varianza .
La problematica più rilevante legata alla costruzione dell' IC 'bias corrected and accelerated ' è sempre dovuta ai tempi di calcolo : per calcolare l' accelerazione occorre calcolare le repliche jackknife dei coefficienti di trasmissione che per questo problema ammontavano a più di due milioni(845 partecipanti Polymod*2446 partecipanti ESEN2) di repliche per ogni q considerato nel modello.
Per utilizzare tecniche di costruzione approssimate ('ABC method')di intervalli di confidenza sarebbe stato necessario considerare le distribuzioni multinomiali bivariate dei profili di seroprevalenza e delle matrici di contatto ; utilizzando quantità pivotali e metodi di approssimazione(delta-method) si sarebbe potuto giungere agli stessi IC calcolati tramite il metodo BCa ,in minor tempo.

9.3.2 Standard error ideali e valutazione della distorsione bootstrap

Nelle tabelle che seguono (Tabella 9.2, Tabella 9.3, Tabella 9.4, Tabella 9.5, Tabella 9.6) sono stati calcolati per i vari modelli stimati M1-M2-M3-M4-M5 i valori dei parametri caratteristici dell'approccio bootstrap ovvero gli standard error, i corrispondenti coefficienti di variazione, la distorsione bootstrap ed il rapporto tra distorsione e se ideale dei parametri di trasmissione. Il calcolo è stato ripetuto per ciascuno dei tre casi considerati,ovvero a seconda che le due fonti di incertezza siano state ricampionate separatamente (colonne 2,3 di ciascuna tabella),oppure congiuntamente(colonna 4 di ciascuna tabella).
Attraverso la stima di questi indici della variabilità è poi possibile distinguere tra IC bootstrap più o meno affidabili per i parametri di trasmissione e valutare appropriatamente la velocità di convergenza allo standard error ideale per ogni modello nei 3 casi considerati.
Nelle tabelle SE indica lo standard error ideale , 'cv' il coefficiente di variazione bootstrap corrispondente allo SE ideale ,'bias' la distorsione bootstrap corrispondente allo SE ideale e 'ratio' il rapporto tra distorsione e standard error ideale.
Il numero di repliche utilizzate per stimare tali valori sono B=P=2000 nel caso in cui gli indici siano condizionati ad una fonte dei dati e (B,P)=(300,300) nel caso in cui la stima di tali indici è ottenuta congiuntamente per le due fonti d'incertezza.Nella Tabella 9.2 sono riportati i valori di convergenza degli indici di variabilità suddetti per il modello M1.

Tabella 9.2:se,cv,bias,ratio=bias/se ideali per il modello M1

Modello:M1 Contatti Totali	Fonte d'incertezza: Sierologia P=2000	Fonte d'incertezza: Contatti B=2000	Fonti d'incertezza: Siero e Contatti (B,P)=(300,300)
SE q	**0.000672**	**0.0043556**	**0.00454**
SE R0	**0.099334**	**0.6435**	**0.6713**
cv q	**0.021011**	**0.13731**	**0.14417**
cv R0	**0.021011**	**0.13731**	**0.14417**
bias q	**0.00004**	**-0.00024**	**-0.000443**
bias R0	**0.0058**	**-0.03543**	**-0.0655**
ratio q	**0.0582**	**-0.05512**	**-0.0976**
ratio R0	**0.0584**	**-0.0550**	**-0.0976**

Per il modello M1, gli standard error ideali condizionati ad una e poi all'altra fonte sono di ordini di grandezza diversi(SE q/contatti=0.000672; SE q/sierologia=0.00435):la fonte che ha un peso maggiore sulla variabilità complessiva (SE q =0.00454) è quella dei contatti .
Sia per il coefficiente di trasmissione q che per l'R_0 la distorsione bootstrap(Tabella 9.2 , righe bias q e bias R0; colonna 4) è trascurabile per un numero di repliche (B,P)=(300,300) ,dato che i valori del rapporto tra bias e standard error (Tabella 9.2 , ratio q e ratio R0,colonna 4) sono inferiori a 0.25 ; inoltre lo se ideale è affidabile dato che il coefficiente di variazione (Tabella 9.2 ;righe cv q e cv R0; colonna 4) è pari al 14.417%.

Tabella 9.3: se,cv,bias,ratio=bias/se ideali per il modello M2

Modello:M2 Contatti per prossimità q1=fisico q2=non fisico	Fonte d'incertezza : Sierologia P=2000	Fonte d'incertezza: Contatti B=2000	Fonti d'incertezza: Siero e Contatti (B,P)=(300,300)
SE q1	**0.000834**	**0.006151**	**0.008166**
SE q2	**0.00000000000094**	**0.001339**	**0.01631**
SE R0	**0.06885**	**0.5584**	**1.3514**
cv q1	**0.01923**	**0.1419**	**0.1853**
cv q2	**35.9084**	**7.8185**	**11.3101**
cv R0	**0.01923**	**0.1533**	**0.3571**
bias q1	**0.00006163**	**0.00002655**	**0.0007542**
bias q2	**0.00000000000002**	**0.0001712**	**0.001442**
bias R0	**0.00509**	**0.06510**	**0.2087**
ratio q1	**0.0739**	**0.004316**	**0.09235**
ratio q2	**0.02608**	**0.1279**	**0.08841**
ratio R0	**0.07392**	**0.1165**	**0.1544**

In Tabella 9.3 sono riportati i principali indici di variabilità del modello M2(contatti per prossimità)
Analizzando il contributo che ognuna delle due fonti dà all'incertezza complessiva del modello M2 appare evidente che il peso maggiore è dato dai dati di contatti sociali .Infatti in Tabella 9.3 gli standard error condizionati ai contatti sociali (Tabella 9.3 ;colonne 2, ;righe SE q1, SE q2, SE R0) hanno sempre ordini di grandezza inferiori agli standard error condizionati ai dati sierologici (Tabella 9.3 ;colonne 3, ;righe SE q1, SE q2, SE R0).La maggiore variabilità potrebbe essere imputata in parte dalla minore numerosità campionaria dei dati di contatto (845 partecipanti) rispetto ai dati di sierologia (2446 partecipanti).
Il parametro q-non fisico (q2) ha una grande variabilità dato che il suo coefficiente di variazione è superiore ad 11(Tabella 9.3 ;colonne 4, ;righe cv q2). La distorsione bootstrap è trascurabile per q1,q2 ed R_0 per il numero di repliche considerato B=P=300, dato che il rapporto tra distorsione e standard error è sempre inferiore a 0.25 (Tabella 9.3 ;colonne 4, ;righe ratio q1, ratio q2, ratio R0).

Tabella 9.4: se,cv,bias,ratio=bias/se ideali per il modello M3

Modello:M3 Contatti per durata q1=c. <15min q2=15<c.<60min q3=c.>60 min	Fonte d'incertezza: Sierologia P=2000	Fonte d'incertezza: Contatti B=2000	Fonti d'incertezza: Sierologia e Contatti (B,P)=(300,300)
SE q1	**0.0006514**	**0.02063**	**0.02219**
SE q2	**0.00852**	**0.05420**	**0.05611**
SE q3	**0.0009178**	**0.01046**	**0.01206**
SE R0	**0.1416**	**0.8042**	**0.8267**
cv q1	**16.05982**	**2.016**	**1.9872**
cv q2	**0.2562**	**1.2911**	**1.2417**
cv q3	**0.02638**	**0.3282**	**0.4030**
cv R0	**0.03538**	**0.1850**	**0.1882**
bias q1	**0.00004056**	**0.01023**	**0.01116**
bias q2	**0.0002937**	**0.009028**	**0.01222**
bias q3	**0.00002298**	**-0.002897**	**-0.004831**
bias R0	**0.01152**	**0.3535**	**0.4011**
ratio q1	**0.06226**	**0.4960**	**0.5032**
ratio q2	**0.03448**	**0.1665**	**0.2179**
ratio q3	**0.025039**	**-0.2768**	**-0.4004**
ratio R0	**0.08135**	**0.4396**	**0.4852**

In Tabella 9.4 sono riportati gli indici di variabilità dei coefficienti di trasmissione per il modello M3 .L'analisi degli standard error condizionati per i parametri q1,q2,q3 (Tabella 9.4;colonne 2,3; righe SE q1,SE q2,SE q3) mostra che ,per il modello M3 ,la fonte dei dati di contatto è sempre quella che porta maggiore variabilità al modello. Per l'analisi congiunta della variabilità(colonna 4),i parametri q1,q2,q3 presentano uno standard error grande dato che il coefficiente di variazione per questi parametri è compreso tra il 40 e 200% (Tabella 9.4;colonna 4;righe cv q1 ,cv q2,cv q3): tale variabilità non si ripercuote sull'R0 che ha un coefficiente di variazione pari a 18,8% .Solo per il parametro q2 la distorsione bootstrap è trascurabile (Tabella 9.4, colonna 4 ,riga 'bias q2'),dato che con P=300 e B=300 , il rapporto tra distorsione e standard error è pari a 0.21(Tabella 9.4, colonna 4 ,riga 'ratio q2').

Per i parametri q1(contatti minori di 15 minuti) ,q3(contatti maggiori di un'ora) ed R0 la distorsione bootstrap al livello di repliche considerato non è trascurabile(Tabella 9.4, colonna 4 ,riga 'ratio q1', 'ratio q3', 'ratio R0') e deve essere considerata (nella costruzione di IC per tali parametri ,si può apportare una correzione per la distorsione).

Tabella 9.5: se,cv,bias,ratio=bias/se ideali per il modello M4

Modello:M4 Contatti per luogo q1=c. casa q2=c. lavoro q3=c. scuola q4=c. altri luoghi	Fonte d'incertezza: Sierologia P=2000	Fonte d'incertezza: Contatti B=2000	Fonti d'incertezza: Sierologia e Contatti (B,P)=(300,100)
SE q1	**0.0002848**	**0.020851**	**0.0225**
SE q2	**0.0000066409**	**0.0028616**	**0.0068**
SE q3	**0.0014146**	**0.011771**	**0.0118**
SE q4	**0.0046271**	**0.018892**	**0.0205**
SE R0	**0.11945**	**0.910596**	**0.9579**
cv q1	**24.430**	**2.160641**	**2.3877**
cv q2	**25.0082**	**20.52286**	**13.4873**
cv q3	**0.03492**	**0.292409**	**0.2930**
cv q4	**0.3586**	**1.468742**	**1.5243**
cv R0	**0.030795**	**0.2238**	**0.2325**
bias q1	**-0.01293**	**-0.003292**	**-0.0035**
bias q2	**0.0000002655**	**0.00013943**	**0.0005**
bias q3	**-0.001506**	**-0.001755**	**-0.0016**
bias q4	**0.012902**	**0.01286**	**0.0134**
bias R0	**0.29274**	**0.4818**	**0.5346**
ratio q1	**-45.4007**	**-0.1578**	**-0.1573**
ratio q2	**0.03998**	**0.048726**	**0.0741**
ratio q3	**-1.064775**	**-0.149161**	**-0.1350**
ratio q4	**2.78832**	**0.680854**	**0.6560**
ratio R0	**2.45074**	**0.5292**	**0.5580**

In Tabella 9.5 sono riportati i valori ideali degli indici di variabilità dei coefficienti di trasmissione del modello M4(contatti per luogo)..
Per il modello M4 la fonte dei dati di contatto è la fonte d'incertezza più rilevante ,per i parametri q1(contatti in famiglia),q2(contatti a lavoro),q4(contatti in altri luoghi) la variabilità della stima bootstrap è elevata (cv q1%=238%, cv q2%=1300%, cv q4%=152%), mentre per q3(contatti a scuola) tale variabilità è contenuta(cv q3%=**29%**) .La distorsione bootstrap ,al livello di repliche considerato ,non è trascurabile per il solo parametro q4(cv q4%=65.6%) : la distorsione di un parametro variabile e significativo si ripercuote sull'R0 rendendo ,a sua volta ,la distorsione per questo parametro non trascurabile (cv R0%=55.8%)).

Tabella 9.6: se,cv,bias,ratio=bias/se ideali per il modello M5

Modello:M5 Contatti per durata q1=c. giornalieri q2=c. settimanali q3=c. sporadici	Fonte d'incertezza: Sierologia P=2000	Fonte d'incertezza: Contatti B=2000	Fonti d'incertezza: Sierologia e Contatti (B,P)=(300,300)
se q1	**0.01633**	**0.0137**	**0.0081**
se q2	**0.1557**	**0.1136**	**0.0469**
se q3	**0.00000011109**	**0.0145**	**0.0328**
se R0	**2.00179**	**1.4698**	**1.0199**
cv q1	**1.7101**	**0.5040**	**0.2322**
cv q2	**0.5854**	**1.4498**	**3.4473**
cv q3	**23.5077**	**0.5677**	**3.0393**
cv R0	**0.26493**	**0.2937**	**0.2297**
bias q1	**-0.02736**	**-0.0098**	**-0.0021**
bias q2	**0.26596**	**0.0784**	**0.0136**
bias q3	**0.00000000472572**	**0.0256**	**0.0108**
bias R0	**3.475**	**0.9241**	**-4.0696**
ratio q1	**-1.6746**	**-0.7183**	**-0.2535**
ratio q2	**1.70804**	**0.6897**	**0.2901**
ratio q3	**0.04253**	**1.7615**	**0.3290**
ratio R0	**1.7360**	**0.6287**	**-3.9901**

Infine,in Tabella 9.6 sono riportati gli indici di variabilità bootstrap per i coefficienti di trasmissione del modello M5(contatti per frequenza).
Per il modello M5 entrambe le fonti d'incertezza sono rilevanti ;non vi è una netta differenza tra l'ordine di grandezza dello se ideale condizionati prima ad una e poi all'altra fonte dei dati(Tabella 9.6;colonna 2,3;righe SE q1,SE q2,SE q3). La maggiore variabilità è quella dei parametri q2(contatti settimanali) e q3(contatti occasionali); i coefficienti di variazione per i due parametri sono rispettivamente del 344% e del 303% (Tabella 9.6;colonna 4;righe ,cv q2 ,cv q3).Per i parametri q1(contatti giornalieri) e per l'R0 i coefficienti di variazione (ideali) sono pari al 23.22% ed al 22.97% rispettivamente : questi due parametri hanno quindi una variabilità contenuta(Tabella 9.6;colonna 4;righe ,cv q1 ,cv R0).
Inoltre per tutti i parametri del modello ,ma soprattutto per l'R_0 la distorsione bootstrap non è trascurabile(Tabella 9.6;colonna 4;righe ,cv q1 ,cv R0), dato che i rapporti tra distorsione e se ideale superano sempre il 'cut off' 0.25 (ratio q1=0.2535; ratio q2=0.2901; ratio q3=0.3290; ratio q4=3.9901).

9.4 Costruzione degli Intervalli di confidenza bootstrap

Di seguito si riportano per i modelli ad informazione esaustivi gli intervalli di confidenza bootstrap(Tabella 9.7,Tabella 9.8,Tabella 9.9,Tabella 9.10,Tabella 9.11) : alcuni di questi intervalli (*)sono corretti per la distorsione bootstrap ,quando quest'ultima non è trascurabile; ciò si verifica quando per alcuni parametri di trasmissione il rapporto tra distorsione e standard error bootstrap(ideale) è superiore alla soglia limite (0.25).

Tabella 9.7:IC 95% bootstrap ,modello M1 contatti totali

M1:Contatti totali	q	R_0
IC normale	**(0.0226 ; 0.0404)**	**(3.3405 ; 5.9720)**
IC percentile	**(0.0232 0.0415)**	**(3.4318 ; 6.1238)**
IC studentized	**(0.0197 ; 0.0394)**	**(2.9172 ; 5.8224)**
IC BC_a	**(0.0217 ; 0.0427)**	**(3.5113 ; 6.3035)**

Tabella 9.8: IC 95% bootstrap modello M2 (contatti per prossimità)

Modello M2	q1	q2	R_0
IC normale	**(0.0281 ; 0.0601)**	**(0 ; 0.0334)**	**(1.1350 ; 6.4327)**
IC percentile	**(0.0319 ; 0.0581)**	**(0.0000 ; 0.0041)**	**(2.6401 ; 5.3502)**
IC studentized	**(0.0308 ; 0.0590)**	**(0 ; 0.0451)**	**(1.5 ; 5.9201)**
IC BC_a	**(0.0286 ; 0.0579)**	**(0.0000 ; 0.03964)**	**(2.5711 ; 4.8106)**

Tabella 9.9: IC 95% bootstrap modello M3(contatti per durata)

Modello M3	q1	q2	q3	R_0
IC normale	**(0;0,0436)***	**(0;0,1552)**	**(0,0113;0.0684)***	**(2,3713;5,60)***
IC percentile	**(0;0,0781)***	**(0,0000;0,2183)**	**(0,0048;0,0549)***	**(2,701;6,040)***
IC studentized	**(0,0000;0,0520)**	**(0,0000;0,2090)**	**(0,0000;0,0408)**	**(2,9031;4,5814)**
IC BC_a	**(0;0,0547)**	**(0,0000;0,2138)**	**(0,0000;0,0560)**	**(2,5116;5,4983)**

* (IC corretti per la distorsione bootstrap)

Tabella 9.10: IC 95% bootstrap modello M4 (contatti per luogo)

Modello M4	q1	q2	q3	q4	R0
IC normale	**(0; 0.0534)**	**(0 ; 0.0139)**	**(0.0172;0.0636)**	**(0; 0.0402)***	**(1.71 ; 5.47)***
IC percentile	**(0, 0.08738)**	**(0;0.000010868)**	**(0.0186 ; 0.0631)**	**(0; 0.067)***	**(2.25 ; 6.04)***
IC studentized	**(0.001; 0.0793)**	**(0;0.00001)**	**(0.0170;0.0650)**	**(0; 0.06123)**	**(2.785 ; 5.678)**
IC BC_a	**(0.00084; 0.0704)**	**(0;0.000012)**	**(0.0164;0.0666)**	**(0; 0.07455)**	**(2.593 ; 6.433)**

* (IC corretti per la distorsione bootstrap)

Tabella 9.11: IC 95% bootstrap modello M5 (contatti per frequenza)

Modello M5	q1	q2	q3	R0
IC normale	**(0,0211 ; 0,0528)***	**(0 ; 0,0920)***	**(0 ; 0.06)***	**(2.0814 6.0794)***
IC percentile	**(0.0167 ; 0.0501)***	**(0.0000 ; 0.17) ***	**(0; 0.1180)***	**(3.0390 ; 7.0387)***
IC studentized	**(0.0253 ; 0.0624)**	**(0;2.87E(-6))**	**(0;4.4E(-26))**	**(1.5; 6.31)**
IC BC_a	**(0.0025 ; 0.0546)**	**(0 ; 0.0023)**	**(0 ; 0.2879E(-18))**	**(2.8304 ; 6.0349)**

* (IC corretti per la distorsione bootstrap)

Conclusioni

Per la varicella in Italia sono state ottenute nuove stime dei coefficienti di trasmissione per una varietà di modelli ed l'individuate le matrici dei contatti sociali che "spiegano" meglio i dati mediante criteri di model selection (AIC,BIC) & inferenza multi-model.
I modelli ad informazione esaustiva hanno un adattamento simile :il modello 'migliore' è quello dei contatti per luogo ,ma si adattano bene anche i modelli dei contatti per prossimità e per frequenza .
Non sembrano rilevanti i contatti di tipo non fisico , quelli nel tempo libero,i contatti occasionali ed i contatti a lavoro.Questo conferma congetture tradizionali per cui le infezioni dei bambini si trasmettono soprattutto a scuola ed a casa. Le stime degli standard error bootstrap delle repliche dei parametri di trasmissione hanno confermato che per questi modelli la maggiore fonte di variabilità è dovuta ai contatti sociali. Questo risultato può essere spiegato sia considerando la minore numerosità campionaria utilizzata nell'indagine Polymod (845 partecipanti italiani) rispetto all'indagine ESEN 2(2446 partecipanti), sia constatando che ciò che differenzia realmente un modello da un altro sono le tipologie di matrici dei contatti scelte ,mentre i profili di seroprevalenza osservati sono gli stessi. In presenza di due fonti di incertezza sono stati calcolati una serie di statistiche bootstrap (se,cv,bias,rapporto tra bias e se) dei parametri di trasmissione capaci di aiutarci nella costruzione degli intervalli di confidenza e nella valutazione del numero minimo di ricampionamenti bootstrap necessari a giungere ai valori 'ideali' delle stime . L'individuazione del numero minimo di repliche bootstrap dei coefficienti di trasmissione mira a ridurre i tempi computazionali di calcolo : l'utilizzo di tecniche di 'nested bootstrap' inevitabilmente si ripercuote sui tempo che occorre per stimare le repliche bootstrap. .Le distribuzioni bootstrap delle repliche dei coefficienti di trasmissione si discostano spesso dall'essere Normali e ciò rende alcuni intervalli di confidenza poco affidabili poiché basati su ipotesi che discordano con l'evidenza empirica data dalle procedure bootstrap.La tecnica bootstrap adottata in questa tesi è non parametrica : non vengono fatte assunzioni sulla distribuzione dei dati : la distribuzione empirica del campione è lo strumento per ricampionare direttamente gli individui che hanno partecipato alle indagine campionarie ESEN 2 e Polymod.Alternativamente si potrebbe usare un bootstrap parametrico , assumendo che i dati di contatto siano distribuiti secondo una binomiale negativa in ogni cella della matrice del numero totale di contatti.
La stratificazione dei contatti si traduce in un progressivo aumento della stocasticità delle matrici di contattoe conseguentemente di una perdita di rilevanza della matrice stessa: in questo caso l'utilizzo del dato grezzo non è opportuno,converrebbe quindi utilizzare un'approccio parametrico per la distribuzione dei contatti.
Dato che si assume che le indagini siano rappresentative della popolazione , gli individui potrebbero essere ricampionati in ogni classe d'età rispettando i pesi campionari del dataset di partenza : ogni campione bootstrap ,in questo modo, risulterebbe rappresentativo della popolazione.
Tale approccio è consistente con il disegno campionario(' *design consistecy')* .
Goeyvaerts et al.(2010) utilizzano un processo di randomizzazione dell'età per costituire i campioni bootstrap : un confronto tra possibili modi di ricampionare risulterebbe utile alla comprensione dei meccanismi che generano la variabilità nei modelli adottati.
In ultimo ,l'utilizzo di metodi bootstrap approssimati ('ABC method') ridurrebbe i tempi computazionali di calcolo .
Tali criticità saranno affrontati in lavori futuri.

Appendice

Function Matlab per calcolare FOI, R0 and likelihood, condizionatamente ai valori parametrici(caso one-q)

```
function q_optimizer=q_optimizer(q_vinc)
global    kappa_matrix  D  ita_total_sero lambda_eq prev_eq ll_model c_eta

size_kappa_matrix=size(kappa_matrix);
size_c_eta=size(c_eta);
class_length_window=zeros(size_c_eta(2)-1,1);
for aux=1:(size_c_eta(2)-1)
    class_length_window(aux,1)=c_eta(1,aux+1)-c_eta(1,aux);
end
class_length_window_days=class_length_window*365 ;
cum_class=cumsum(class_length_window);
D=7 ;
lambda_0=(1/365)*0.1*ones(size_kappa_matrix(1),1);

tol=1;
it=1;
x_a=zeros(size_kappa_matrix(1),1);
delta=zeros(size_kappa_matrix(1),1);
differenza =zeros(size_kappa_matrix(1),1);
foi=lambda_0;

while(tol>5*10^(-30) && it<=5*10^(5))

x_a(1)=exp(-foi(1)*class_length_window_days(1));
for j=2:(size_kappa_matrix(1))
x_a(j)=x_a(j-1)*exp(-foi(j)*class_length_window_days(j));
end
delta(1)=(1-exp(-foi(1)*class_length_window_days(1)));
for j=2:(size_kappa_matrix(1))
delta(j)=x_a(j-1)*(1-exp(-foi(j)*class_length_window_days(j)));
end
foi_next=exp(q_vinc)*D*kappa_matrix*delta;
differenza=foi-foi_next;
tol=sum((foi_next-foi).^2);
it=it+1;
foi=foi_next;
end
if it==5*10^(5)
error('Maximum number of iterations exceeded')
end
```

……(continua)……..

Bibliografia

Anderson, R. M. and R. M. May (1991). Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press.

Beutels P, Shkedy Z, Aerts M, van Damme P. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a Web-based interface. Epidemiol Infection 2003; 134 (6): 1158–1166.

Bradley Efron and Robert J. Tibshirani(1993),An introduction to the bootstrap. Monographs on Statistics and Apllied Probabilità.Chapman and Hall .

Burnham, K. P. and D. R. Anderson (2002). Model Selection and Multi-Model Inference:A Practical Information-Theoretic Approach. Springer-Verlag New York Inc.

Cox and Hinkley (1974), Theoretical Statistics,Chapman and Hall

Davison and Hinkley (1997), Bootstrap methods and thieir applications

Diekmann, O., J. A. P. Heesterbeek, and J. A. J. Metz (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology 28,65–382.

Edmunds WJ, Kafatos G, Wallinga J, Mossong J. Mixing patterns and the spread of close-contact infectious diseases, Emerg Themes Epidemiol. 2006; 3: 10.

Edmunds WJ, O'Callaghan CJ, Nokes DJ (1997) Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proceedings of the Royal Society of London B 264: 949-957.

Farrington, C. P. , M. N. Kanaan, and N. J. Gay (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. Applied Statistics 50,251–292.

Farrington, C. P. and H. J. Whitaker (2005). Contact surface models for infectious diseases: estimation from serologic survey data. Journal of the American Statistical Association 100,370–379.

Garnett, G. P. and B. T. Grenfell (1992). The epidemiology of varicella-zoster virus infections: a mathematical model. Epidemiology and Infection 108,495–511.

Goeyvaerts N., Niel Hens, Benson Ogunjimi, Marc Aerts, Ziv Shkedy, PierreVan Damme, Philippe Beutels(2010), Estimating infectious disease parameters from data on social contacts and serological status

Grenfell B. T. and Anderson R. M. (1985),The estimation of age-related rates of infection from case notifications and serological data.Department of Pure and Applied Biology ,Imperial College,London University.

Greenhalgh, D. and K. Dietz (1994). Some bounds on estimates for reproductive ratio derived from the age-specific force of infection. Mathematical Biosciences 124,9–57.

Hethcote Herbert W. (1996),Modeling Heterogeneous Mixing in Infectious Disease Dynamics. Models for infectious human diseases,Their Structure and Relation to the Data.Cambrige University

Hethcote Herbert W. (1989),Three basic epidemiological models. BioMathematics,Vol. 18,Springer-Verlag,Berlin.

Lagarias, J.C., J. A. Reeds, M. H. Wright, and P. E. Wright(1990), "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM Journal of Optimization

Melegaro A.,Mark Jit, Nigel Gaya,Emilio Zagheni, W. John Edmunds(2007),What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns (‘Unpublished paper’).

Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, et al. (2008b). Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Medicine 5(3), 10.1371/journal.pmed.0050074.

Shkedy Z., M. Aerts, G. Molenberghs, P. Van Damme,and P.Beutels(2006).Modelling age-dependent force of infection from prevalence data usung fractional polynomials. Statistics in Medicine ,Wiley InterScience(www.interscience.wiley.com)

Van Effelterre, T. , Z. Shkedy, M. Aerts, G. Molenberghs, P. Van Damme,and P.Beutels (2009). Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. Epidemiology and Infection 137,48–57.

Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am J Epidemiology 2006; 164(10):936–944.

Whitaker, H. J. and C. P. Farrington (2004). Infections with varying contact rates: application to varicella. Biometrics 60,615–623.

Wood, S. N. (2006). Generalized Additive Models: an Introduction with R.Chapmanand Hall/CRC Press.

Zagheni E.(2008),Using Time Use Data to Parameterize Models for the Spread of Close-contact Infectious Diseases.