Lo studio della nuzialità con dati di censimento:

dalla SMAM all'*Event History Analysis* (e ritorno)

ALESSANDRO ROSINA

1. Introduzione. I metodi di *survival analysis* (ma più in generale di *event history analysis*) sono stati sviluppati allo scopo di poter analizzare adeguatamente la durata di attesa di un evento di interesse. L'esigenza di analizzare dati di tale tipo è comune a molte discipline (medicina, biologia, ingegneria, economia, sociologia, demografia, ecc.). L'analisi della mortalità (della sopravvivenza fino all'evento morte) è un esempio tipico. Ma allo stesso modo è possibile studiare tutti gli altri eventi biografici (il matrimonio, la nascita di un figlio, ecc.).

È comune nei dati di durata la presenza di osservazioni incomplete. Un'osservazione viene detta completa (non censurata) quando è nota con precisione la data sia dell'evento origine[1] che dell'evento di interesse[2], e di conseguenza è nota la durata di attesa di sperimentazione dell'evento di interesse (durata di sopravvivenza nel caso della mortalità, durata del matrimonio nel caso della divorzialità, ecc.).

Ci sono sostanzialmente due tipi di censura: a destra e a sinistra[3]. La censura a destra si ottiene quando il periodo di osservazione dell'evento si interrompe prima che accada l'evento di interesse. Supponiamo di studiare in un dato paese la mortalità di una data generazione. Un individuo che all'età x emigra costituisce una osservazione incompleta, ed in particolare censurata a destra. Noi sappiamo infatti solamente che fino all'età x tale individuo era vivo ma non sappiamo da x in poi quando morirà. Ci si trova invece nel caso di censura a sinistra quando è noto che l'evento è accaduto ma non si sa quando. Ad esempio in un'indagine retrospettiva «to determine the distribution of the time until first marijuana use among high school boys in California [...] the question was asked "When did you first use marijuana?". One of the responses was "I have used it but can not recall just when the first time was"» (Klein, Moeschberger 1997).

I dati *current-status* sono una forma particolare di dati censurati che corrisponde all'estrema situazione nella quale nessuna osservazione è completa (ovvero per nessun individuo si conosce la data di sperimentazione dell'evento di interesse), si dispone infatti di osservazioni che sono o censurate a destra (si sa che fino al momento della rilevazione l'individuo non aveva ancora sperimentato l'evento) o censurate a sinistra (si sa che l'individuo ha sperimentato l'evento in un non precisato momento precedente la rilevazione).

Ci si può trovare in tale situazione tipicamente quando in una rilevazione trasversale (come nel caso di un censimento) si dispone solamente dell'informazione sul fatto che la persona abbia o meno già sperimentato l'evento di interesse.

SIDeS, «Popolazione e Storia»,
2/2005, pp. 67-78

A volte, anche se l'informazione sulla data dell'evento è disponibile si preferisce non utilizzarla perché non sufficientemente affidabile. «Current-status data generally are considered more reliable then retrospective reports of age or duration, because respondents can more accurately report their current state than recall the time at which some event occured in the past» (Mansmann 2000). Ciò è vero in particolare per alcuni tipi di eventi non facilmente collocabili nel tempo da chi li ha vissuti, e soprattutto per le rilevazioni nei paesi in via di sviluppo e per i censimenti e le rilevazioni in epoca storica. Secondo Diamond *et al.* (1993) ciò può inoltre valere anche per dati ottenuti da ricostruzioni nominative: «where there is some doubt over a particular group of record linkages – caused perhaps by the records being prepared by different scribes who used different abbreviations – then the use of current-status data may offer a strategy for estimating accurately the effects of some covariate».

Il caso più comune in demografia storica di uso di dati di questo tipo è quello della stima dell'età media al primo matrimonio e del nubilato/celibato definitivo a partire da dati di censimento o da stati delle anime. Il contributo seminale è quello di Hajnal che mezzo secolo fa proponeva una stima non parametrica della funzione di sopravvivenza in stato di celibe/nubile e illustrava come ottenere da tale funzione l'età media al matrimonio. Nel paragrafo 2 rivisiteremo tale procedura nell'ottica dei concetti e delle notazioni proprie delle tecniche non parametriche di *event history analysis*.

L'estensione dei modelli di *event history analysis* per l'uso di dati *current status* è stata inizialmente proposta da Diamond, McDonald e Shah nel 1986. Tali autori conducono la propria analisi sull'età allo svezzamento. L'interesse per l'applicazione dei modelli di *event history analysis* per dati *current status* si estende in molte altre discipline. In campo epidemiologico un esempio tipico è l'infezione da HIV (si ha l'informazione su chi è infetto ma difficilmente è nota in modo affidabile la data di quando l'infezione è stata contratta; Jewell, Shiboski 1990), ed in generale l'insorgenza di una data malattia (spesso si hanno accurati test diagnostici sulla presenza della malattia in un individuo, ma non si è in grado di stabilire quando è insorta).

L'obiettivo rimane comunque quello dei modelli di *event history analysis*, ovvero lo studio dinamico del processo che produce un dato evento di interesse in funzione di opportuni fattori esplicativi. Come vedremo nel terzo paragrafo, la specificità dei dati *current-status*, in particolare la mancanza di informazione sulla data dell'evento, impone però delle soluzioni *ad hoc*.

La differenza maggiore rispetto ai classici modelli di analisi della sopravvivenza, è il fatto che ad essere modellata non è direttamente la funzione di rischio (che esprime il rischio di subire l'evento in un dato istante t[4] condizionatamente a non averlo sperimentato fino a quel momento) ma la funzione di ripartizione (che esprime la probabilità di aver già sperimentato l'evento al tempo t). Ciò ha inoltre come conseguenza il fatto di dover vincolare la funzione di ripartizione ad essere monotona crescente (mentre la funzione di rischio deve essere semplicemente non negativa). Introdurremo nel quarto paragrafo alcuni modelli facilmente stimabili con i

pacchetti statistici più comuni. Proporremo inoltre un'applicazione dimostrativa ai dati di Oderzo desunti da uno status animarum del 1672[5]. Verrà discussa infine una possibile estensione, per tener conto del fatto che il matrimonio è un evento evitabile, e quindi che una quota della popolazione è destinata a non sperimentare l'evento di interesse.

2. Una rilettura della SMAM. Come sottolineano Del Panta e Rettaroli (1994) «l'analisi demografica del matrimonio mira a rispondere ad alcuni quesiti fondamentali. In primo luogo occorre sapere quale sia il livello e cioè l'intensità del fenomeno nuziale. In termini quantitativi ciò significa conoscere la proporzione di coloro che accedono al matrimonio, e quindi, per differenza, la parte di coloro che ne restano esclusi. In secondo luogo si vorrà conoscere come i matrimoni si distribuiscono secondo le età degli sposi, e quale sia quindi la cadenza della nuzialità; se la distribuzione rivela una propensione a sposarsi precoce o tardiva; quale sia l'età in cui mediamente uomini e donne si sposano».

Avendo a disposizioni dati di stato che riportano la composizione per sesso, età e stato civile della popolazione una misura dell'intensità finale della nuzialità viene usualmente ottenuta dal complementare della proporzione di celibi/nubili a 50 anni, mentre l'età media al matrimonio viene fornita dalla tecnica proposta da Hajnal (1953), e nota con l'acronimo di SMAM (Singulate Mean Age at Marriage).

Se la rilevazione trasversale (censimento o stato delle anime) invece di fornire solo lo stato coniugale delle persone riportasse anche l'età al primo matrimonio, saremmo nella situazione classica di un'informazione retrospettiva sulla data dell'evento di interesse, che consentirebbe analisi ordinarie di tipo event history. Mancando la data e disponendo solo dello stato coniugale al momento della rilevazione siamo nella situazione di dati *current status*.

Come è noto (Del Panta, Rettaroli 1994; Hinde 1998; Bonarini, Rosina 2000) la SMAM è una stima corretta dell'età media al matrimonio se il processo di nuzialità rimane sostanzialmente invariato tra la generazione che ha 15-19 anni al momento della rilevazione e quella che ne ha 45-49; inoltre deve essere trascurabile l'effetto selezione della mortalità e delle migrazioni.

Se valgono infatti tali condizioni il rapporto nubili in una data età sulla popolazione totale di tale età consente di ottenere una stima non parametrica della funzione di sopravvivenza nello stato di nubile (S_x). Il complementare ($F_x = 1 - S_x$) fornisce la funzione di ripartizione della primo-nuzialità (probabilità di non essere nubili all'età x, ovvero di arrivare già sposati all'età x).

A partire da tali distribuzioni non è possibile calcolare l'età media al matrimonio, bisogna infatti ricondursi alle funzioni condizionate all'aver sperimentato l'evento (Bonarini, Rosina 2000).

Indichiamo allora con Y la variabile che distingue la popolazione complessiva in 'nubili definitive' (Y = 0) e in 'coniugate definitive' (Y = 1). Utilizzando il linguaggio adottato nei mixture models (Maller, Zhou 1996) chiamiamo 'long-term survivors' le donne del primo gruppo e 'suscettibili' le donne del secondo gruppo.

Sia inoltre: $Pr(Y = 1) = p$; $Pr(Y = 0) = 1 - p = s$.

La funzione di sopravvivenza complessiva (detta 'marginale': S_x) può essere ottenuta come media ponderata della funzione di sopravvivenza delle 'suscettibili' al matrimonio ($S_{x|Y=1}$) e delle 'nubili definitive' ($S_{x|Y=0}$):

$$S_x = p\, S_{x|Y=1} + (1\text{-}p)\, S_{x|Y=0} = p\, S_{x|Y=1} + s$$

Infatti $S_{x|Y=0} = 1$, dato che le nubili definitive non sperimenteranno mai l'evento.
Vale inoltre: $F_x = p\, F_{x|Y=1}$
Dalla relazione precedente si ricava:

$$S_{x|Y=1} = (S_x - s) / p = (S_x - s) / (1 - s)$$

La funzione di sopravvivenza dei suscettibili è quindi direttamente ottenibile dalla funzione di sopravvivenza complessiva, se è nota la quota 's' (o il complementare 'p') di persone che destinate ad evitare definitivamente l'evento (ovvero la quota di nubilato definitivo).

Come è noto l'età media può essere ottenuta dall'integrale (la somma nel discreto) della funzione di sopravvivenza (ad esempio la vita media è ottenuta dalla somma degli L_x della tavola di mortalità, per l_0=1); si ottiene allora infine:

$$\bar{x} = \sum S_{x|Y=1} = \frac{\sum [S_x - s]}{p}$$

che corrisponde (considerando la funzione di sopravvivenza pari a 1 fino ai 15 anni e stimando s con la proporzione di nubili a 50 anni) alla formula della SMAM.

3. Modelli di *event history analysis* per dati *current-status*. Indichiamo con X la variabile 'età al primo matrimonio'. Nella situazione usuale di dati event history da osservazione retrospettiva noi conosciamo il valore di tale variabile per una parte degli individui (osservazione completa) mentre per gli altri individui l'osservazione è censurata a destra (ovvero sappiamo solo che fino all'età che avevano al momento della rilevazione non avevano ancora sperimentato l'evento). Nel caso di dati *current status* quello che cambia è il fatto che gli individui con osservazione completa ora sono censurati a sinistra (ossia, sappiamo solo che hanno sperimentato l'evento precedentemente alla rilevazione, ma non sappiamo quando).

Indichiamo con C la variabile che indica lo stato coniugale al momento della rilevazione. Sia C = 0 per le osservazioni censurate a destra (evento non ancora sperimentato: ancora nubili all'età della rilevazione) e C = 1 le censure a sinistra (evento già accaduto: già sposate all'età della rilevazione).

Per la generica donna *i* noi non conosciamo il valore dell'età al primo matrimonio (X) ma disponiamo solo del valore di C_i e dell'età al momento della rilevazione (A).

Vale quindi:

$$Pr(C = 1|A = a) = Pr(X \leq a) = F(a)$$

dove *a* è il generico valore assunto dalla variabile A.

Ad esempio $Pr(C = 1|A = 25) = Pr(X \leq 25) = F(25)$ è la probabilità di sposarsi

entro i 25 anni (che corrisponde alla probabilità di trovare già sposata una donna che al momento della rilevazione aveva 25 anni).

Se si assume l'indipendenza tra A ed X (ed è ragionevole farlo dato che il momento della rilevazione non è usualmente legato alla propensione al matrimonio nella popolazione studiata), si dimostra che la stima della funzione di ripartizione F(.) può essere ottenuta dalla verosimiglianza condizionata di C (Jewell, van der Laan 2002).

Questo è particolarmente interessante, dato che i parametri del modello di regressione sulla variabile di durata X (non osservata) equivalgono ai parametri del modello con variabile dipendente C (osservata). Di fatto ci si riconduce ad un modello a risposta binaria.

In termini generali, se si pone:

$$\pi = \Pr(C = 1|A = a)$$

possiamo scrivere (Huang, Wellner 1996):

$$g(\pi) = \nu_0(a) + \beta_z z$$

dove g(.) è uno specifico *link* tra variabile dipendente e fattori esplicativi; $\nu_0(.)$ è una funzione monotona non decrescente; z rappresenta i fattori esplicativi e β i corrispondenti parametri.

Ciò corrisponde ad un modello lineare generalizzato (Doksum, Gasko 1990), che può essere stimato con qualsiasi software standard per modelli a risposta binaria.

Se si adotta $g(\pi) = \text{logit}(\pi)$ si ottiene un classico modello di regressione logistica. Se invece si adotta $g(\pi) = \log(-\log(1-\pi))$ si ottiene il noto modello *complementary log-log*, alla cui base, analogamente al modello di Cox, sta l'ipotesi di rischi proporzionali.

3.1. Un modello a rischi proporzionali con età in classi. Diamond *et al.* (1993) hanno illustrato per i demografi storici, nel volume *Old and New Methods in Historical Demography*, come applicare, in modo semplice, un modello a rischi proporzionali su dati *current status*.

Gli eventi vengono raggruppati in intervalli temporali mutamente esclusivi ed esaustivi, con limiti: $0 = a_0 < a_1 < \ldots < a_N$.

La variabile dipendente è:

$$\Pi_j = \Pr(0 \text{ ″ } X < a_j), \text{ dove } j = 1,2,\ldots,N-1$$

La trasformata *complementary log-log* di Π_j consente di pervenire ad un modello a rischi proporzionali:

$$\log(-\log(1- \Pi_j)) = c_j + \beta_Z Z$$

dove: c_j corrisponde al logaritmo della cumulata (fino alla classe di età j) della funzione di rischio di base (baseline function); Z è la generica variabile esplicativa; β_Z è il parametro che ne esprime l'effetto.

Così specificato il modello è ricondotto ad una semplice regressione su dati binari e può quindi essere stimato con i *software* statistici più diffusi.

Come analisi illustrativa, applichiamo tale modello ai dati dello stato delle anime

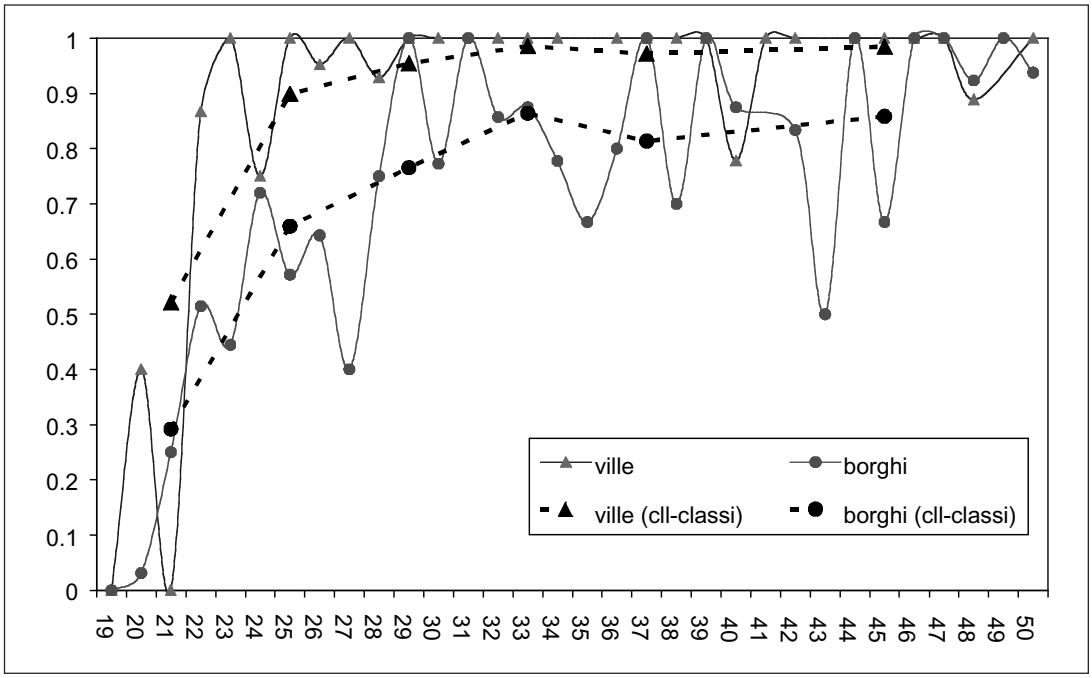
di Oderzo nel 1672. L'unica variabile esplicativa che consideriamo corrisponde alla distinzione tra località denominate 'borgo' e località denominate 'ville'. L'uso di una sola covariata è una semplificazione, giustificata qui dai fini illustrativi dell'applicazione proposta. Tipicamente le variabili esplicative potranno essere più di una, nella logica della regressione multipla. Le classi di età considerate hanno come limiti 22, 26, 30, 34, 38, 50. Dato che nessuna delle ragazze di età fino ai 18 risulta sposata consideriamo tale età come inizio del processo di nuzialità.

I risultati sono riportati in tabella 1 e figura 1[6]. Come si vede, la funzione di ripartizione stimata dal modello sembra interpolare abbastanza bene i valori della proporzione di già sposate per età (in anni). La variabile esplicativa ha effetto significativo negativo indicando un rischio di accesso al matrimonio più basso nei borghi rispetto alle ville[7].

Tab. 1. *Stime del modello complementary log-log con età in classi. Oderzo 1672*

	Parametro	Standard error	p-value
a_{22}	-0.3075	0.1831	0.0931
a_{26}	0.8289	0.1516	<.0001
a_{30}	1.1270	0.1618	<.0001
a_{34}	1.4458	0.2252	<.0001
a_{38}	1.2741	0.2169	<.0001
a_{50}	1.4259	0.1629	<.0001
β_z	-0.7556	0.1344	<.0001

Fig. 1. *Proporzione di donne uscite dallo stato di nubile. Stime dirette e stime dal modello complementary log-log con età in classi. Oderzo 1672*



Il limite principale di questo modello è che non rispetta la condizione che la funzione di ripartizione sia monotona non crescente. Ad esempio, nella nostra applicazione la stima di 'a_{34}' presenta un valore superiore rispetto ad 'a_{38}' e ad 'a_{50}'. Inoltre le classi sono arbitrarie. Infine il modello è poco parsimonioso: un modello parametrico o non parametrico con *splines* potrebbe ottenere uno stesso adattamento con minori parametri.

3.2. Una parametrizzazione del rischio: il modello Weibull. In modo più generale il modello a rischi proporzionali potrebbe essere scritto nel seguente modo:

$$\log(-\log(1- \pi)) = \beta_0(a) + \beta_z Z$$

dove $\beta_0(a)$ rappresenta genericamente la cumulata del rischio di base (baseline function) e $\beta_z Z$ la dipendenza da covariate.

Se si sceglie la distribuzione Weibull per il rischio di base, caratterizzata da $F(a) = 1 – \exp(-\lambda a^b)$, si ottiene il seguente modello:

$$\log(-\log(1- \pi)) = b \log(a) + \beta + \beta_z Z$$

dove $\beta = \log(\lambda)$.

Si tratta di un modello altrettanto semplice da stimare, ma più rigoroso ed elegante rispetto al precedente (non richiede la costruzione di arbitrarie classi di età e consente di ottenere una funzione di ripartizione monotona crescente). Anche questo modello può essere infatti stimato con lo stesso software usato per il modello precedente. I parametri β e b stimati possono poi essere utilizzati per ricavare le funzioni della distribuzione Weibull.

In tabella 2 riportiamo le stime del modello applicato ai dati di Oderzo, ed in figura 2 la funzione di ripartizione corrispondente. Come si vede i parametri stimati sono ora molto meno rispetto al modello precedente e la funzione F(.) risulta monotona crescente.

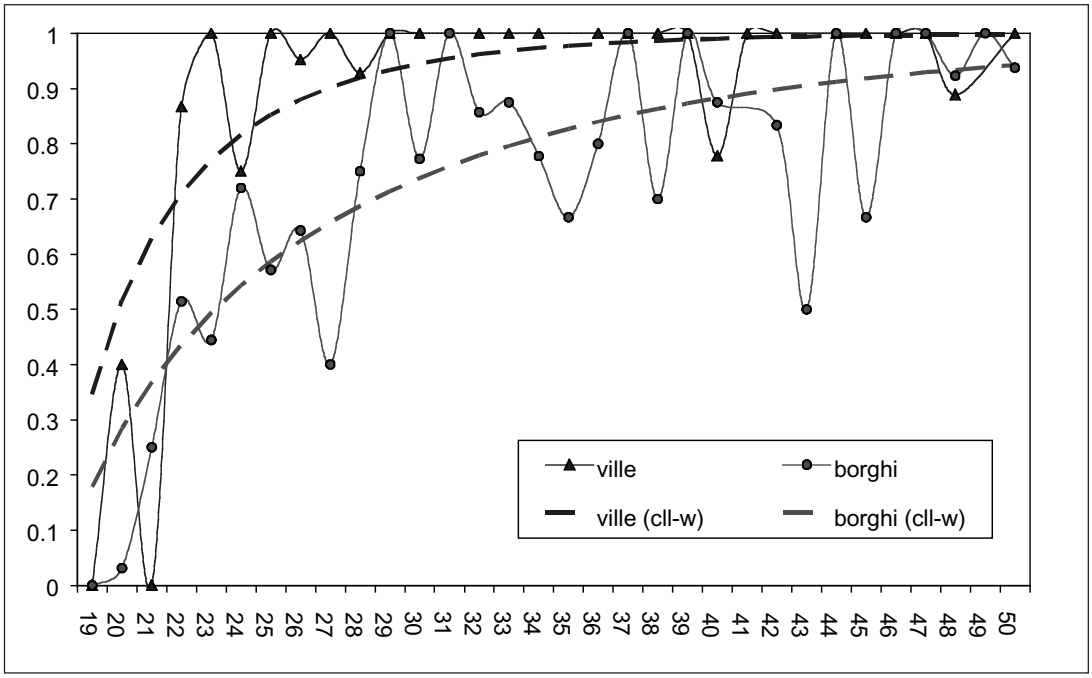
Il limite maggiore è però ora dovuto al fatto che la funzione di ripartizione continua tendenzialmente a crescere fino a raggiungere l'unità. Il che corrisponde all'ipotesi che prima o poi (estrapolando la curva) tutti si sposino. Ovvero l'effetto delle variabili esplicative (β_z) a rigore va letto tutto in termini di cadenza, visto che l'intensità finale (quota di persone che alla fine si sposano) è sempre pari a 1. Per definizione infatti F(a) è non solo monotona crescente ma vale anche $F(a) = 1$ per $x \to \infty$.

3.3. Una proposta di estensione (con ritorno alla logica della SMAM). I modelli considerati (quelli più usati in letteratura con dati *current-status*) non tengono esplicita-

Tab. 2. *Stime del modello Weibull. Oderzo 1672*

	Parametro	Standard error	p-value
b	-0.8561	0.2093	<.0001
β	0.7731	0.0845	<.0001
β_z	-0.7738	0.1356	<.0001

Fig. 2. *Proporzione di donne uscite dallo stato di nubile. Stime dirette e stime dal modello Weibull. Oderzo 1672*



mente conto del fatto che il matrimonio è un evento evitabile, e quindi che una parte della popolazione è destinata a non sperimentare mai l'evento di interesse.

I modelli di 'mistura' (Farewell 1982; Maller, Zhou 1996; Rosina 2000; McDonald, Rosina 2002) costituiscono una generalizzazione dei modelli di *event history analysis* che consente di tener conto non solo del tempo di accadimento dell'evento, ma anche, esplicitamente[8], della possibilità di evitarlo. Tale estensione risponde contemporaneamente a due esigenze: quella tecnico-statistica di poter consentire che la funzione di ripartizione non debba necessariamente raggiungere l'unità (ovvero che possa essere una funzione impropria), in modo quindi di adattarsi meglio al fenomeno di studio, e l'obiettivo sostanziale di consentire di distinguere l'effetto dei fattori esplicativi nell'impatto sull'intensità finale e quello sulla cadenza.

Per introdurre tale estensione anche nel caso di dati *current status*, può essere utile ripartire dalla logica alla base della costruzione della SMAM.

Nel secondo paragrafo, rileggendo la SMAM, avevamo indicato con Y = 0 le 'nubili definitive' (Y = 0) e con Y = 1 le '(già) coniugate definitive' (Y = 1). Con Pr(Y = 1) = p; Pr(Y = 0) = 1 – p = s.

Avevamo visto inoltre che valevano le seguenti relazioni (ci mettiamo qui però in ambito di tempo continuo):

$$S(x) = p\, S(x|Y = 1) + s;\ F(x) = p\, F(x|Y = 1)$$

dove S(x|Y = 1) e F(x|Y = 1) sono rispettivamente la funzione di sopravvivenza e la funzione di ripartizione condizionatamente all'essere destinati a sperimentare prima o poi l'evento, ovvero al fatto che alla fine ci si sposi (Y = 1).

F(x|Y = 1) è quindi una funzione propria dato che è riferita solamente a coloro che sono destinati a sperimentare l'evento:

$$\lim_{x\to\infty} F(x \mid Y = 1) = 1$$

mentre F(x), che invece rappresenta il processo complessivo (di tutta la popolazione), sarà tipicamente impropria:

$$\lim_{x\to\infty} F(x) = p \cdot \lim_{x\to\infty} F(x \mid Y = 1) = p \leq 1$$

La relazione chiave diventa quindi la seguente scomposizione:

$$F(x) = p\, F(x|Y = 1)$$

dove p rappresenta l'intensità finale, mentre F(x|Y = 1) rappresenta la cadenza, da cui ricavare eventualmente una misura di cadenza media. Abbiamo infatti visto nel paragrafo 2 come la SMAM si ottenga da una sintesi di F(x|Y = 1), ovvero di S(x|Y = 1) = 1 - F(x|Y = 1).

Si tratta quindi di costruire non una unica equazione di regressione, bensì due equazioni interdipendenti, ad esempio del tipo:

$$\text{logit}(p) = \alpha_0 + \alpha_z Z$$
$$\log(-\log(1 - F(x|Y = 1))) = \beta_0(x) + \beta_z Z$$

La prima equazione consente di stimare l'effetto di Z su p = Pr(Y = 1), ovvero sulla probabilità di essere una '(già) coniugata definitiva', mentre la seconda equazione permette di stimare l'effetto di Z sull'età al primo matrimonio condizionatamente all'essere una '(già) coniugata definitiva'.

Da notare che la variabile Y è parzialmente latente, ovvero ha valore pari a 1 per le censurate a sinistra (sappiamo che prima della rilevazione si sono sposate) ma ha valore non noto (mancante) per le censurate a destra. Infatti, dato che al momento della rilevazione la generica donna non era ancora sposata non sappiamo se poi si sposerà (Y = 1) o rimarrà definitivamente nubile (Y = 0). È per tale motivo che tali due equazioni sono interdipendenti. La seconda equazione è infatti condizionata alla prima, ma il valore di Y risente anche di come varia il rischio in funzione dell'età (seconda equazione).

Il modello diventa però più complicato ed è necessario ricorrere a tecniche di stima non standard, come ad esempio i metodi MCMC (McDonald, Rosina 2001). Uno sviluppo metodologico in questa direzione si trova in Rosina (2006).

4. Considerazioni conclusive. I dati *current-status* sono una forma particolare di dati censurati che corrisponde all'estrema situazione nella quale nessuna osservazione è completa, si dispone infatti di osservazioni che sono o censurate a destra (si sa che fino al momento della rilevazione l'individuo non aveva ancora sperimentato l'evento) o censurate a sinistra (si sa che l'individuo ha sperimentato l'evento in un non precisato momento precedente la rilevazione). Ci si può trovare in tale situazione tipicamente quando in una rilevazione trasversale (come nel caso di un censimento) si dispone solamente dell'informazione sul fatto che la persona abbia o

meno già sperimentato l'evento di interesse. Si tratta di una condizione molto comune in demografia storica, nei paesi in via di sviluppo, ma anche negli attuali paesi occidentali può spesso accadere che per vari motivi la data dell'evento di interesse non sia disponibile o non venga ritenuta sufficientemente affidabile.

La demografia storica conosce da mezzo secolo un metodo per calcolare in modo non-parametrico l'età media al primo matrimonio a partire da dati *current-status*. Si tratta della SMAM, proposta nel 1953 da Hajnal. Recentemente si è assistito nella letteratura metodologico-statistica ad un rinnovato interesse per tali dati, in particolare come conseguenza dello sviluppo dei modelli di *event history analysis* e dell'esigenza di studiare in epidemiologia le determinanti dell'insorgenza di specifiche malattie senza avere l'informazione di quando la malattia è stata contratta. Ciò ha portato negli ultimi anni a varie proposte di estensione della metodologia dell'*event history analysis* nell'ambito delle applicazioni con dati *current-status* (Jewell, van der Laan 2002). In questo lavoro abbiamo presentato due semplici modelli di base (modello a rischi proporzionali e modello Weibull) che hanno il vantaggio di poter essere stimati utilizzando procedure standard per analisi di modelli a risposta binaria (presenti nei più diffusi pacchetti statistici). Un limite dei modelli finora proposti in letteratura deriva dal fatto che non tengono esplicitamente conto della possibilità che una quota rilevante della popolazione possa sottrarsi all'evento di studio. Una estensione in tale direzione consentirebbe infatti di ricondurre l'analisi con modelli *event history* alla logica originale della SMAM, che invece prevede esplicitamente la presenza di 'long-term survivors', ovvero di nubili definitive (nel caso il processo di studio sia la nuzialità).

Estensioni dei modelli di *event history analysis* con *long-term survivors* (chiamati 'modelli di mistura') su dati standard di durata sono già consolidate in letteratura (Maller, Zhou 1996), si tratta quindi di sviluppare estensioni di tale approccio nel caso di dati *current-status*. Del resto, come abbiamo dimostrato in questo lavoro, la SMAM può essere considerata in modo naturale come una stima non-parametrica all'interno della logica dei modelli di mistura. Il vantaggio dell'estensione dei modelli di *event history analysis* con *long-term surivors* per dati *current-status* consiste nel fatto che il modello diventa in grado di differenziare l'effetto di una variabile esplicativa nel diverso impatto sull'intensità finale e sulla cadenza. Consentendo quindi di distinguere tra situazioni in cui l'effetto agisce solamente sull'intensità finale o solo sulla cadenza, o su entrambe[9].

[1] Ad esempio la data di nascita nel caso di analisi della mortalità; la data di matrimonio nel caso di analisi della divorzialità, ecc.

[2] Ad esempio la data di morte nel caso di studi sulla mortalità; la data di divorzio nel di studi sulla divorzialità, ecc.

[3] Per un approfondimento sui vari sottocasi di osservazioni censurate si rimanda a Yamaguchi (1991).

[4] Il tempo è nei modelli di *event history analysis* inteso come la durata di attesa dell'evento di interesse. Nel caso della nuzialità si tratta semplicemente dell'età (eventualmente a partire da dall'età minima al matrimonio).

[5] L'autore ringrazia Fiorenzo Rossi per aver gentilmente fornito i dati qui utilizzati.

[6] Per la stima abbiamo utilizzato il pacchetto statistico SAS System, ed in particolare la pro-

cedura *genmod* (specificando come link la complementary log-log e come distribuzione la binomiale), ma qualunque software per la stima di modelli a risposta binaria sarebbe stato ugualmente appropriato.

[7] Facciamo anche notare che è possibile inoltre ricavare il nubilato definitivo a partire dai parametri stimati dal modello:
per le ville si ottiene:
$\pi_{50|ville} = 1 - \exp(-\exp(1.4259)) = 0.98$
per i borghi:
$\pi_{50|ville} = 1 - \exp(-\exp(1.4259 - 0.7556)) = 0.86$
Quindi il nubilato definitivo risulta essere pari al 2% nelle ville e al 14% nei borghi.

[8] Alcuni modelli contemplano implicitamente che l'evento studiato possa non verificarsi. Ad esempio il modello a rischi proporzionali di Cox è adeguato anche per l'analisi di processi in cui rimane una frazione di sopravviventi finale.

[9] Ad esempio, nell'applicazione esemplificativa trattata nel testo, 'ville' e 'borghi' potrebbero aver avuto la stessa intensità finale e le differenze essere solo da attribuire ai diversi tempi di accesso al matrimonio. O viceversa i tempi di accesso al matrimonio essere gli stessi e l'intensità finale diversa. Inoltre, in una situazione nella quale una variabile ha un effetto opposto su intensità finale e cadenza (ad esempio, rispetto alla categoria di riferimento, potrebbe essere maggiore la quota di chi non si sposa mentre chi si sposa lo fa in età molto più giovane) con i modelli standard si può ottenere un effetto (complessivo) nullo. Mentre con un modello di mistura l'azione opposta della variabile su intensità finale e cadenza viene adeguatamente colta.

Riferimenti bibliografici

F. Bonarini, A. Rosina 2000, *Appunti di demografia. Parte II: Nuzialità e formazione delle unioni*, CLEUP, Padova.

L. Del Panta, R. Rettaroli 1994, *Introduzione alla demografia storica*, Laterza, Roma-Bari.

I. Diamond, J.W. McDonald, I.H. Shah 1986, *Proportional hazard models for current status data: application to the study of differentials in age at weaning in Pakistan*, «Demography», 23, 607-620.

I. Diamond, J.W. McDonald 1991, *The analysis of current status data*, in J. Trussel, R. Hankinson and J. Tilton (eds.), *Demographic Application of Event History Analysis*, Oxford, Oxford University Press.

I. Diamond, D. Rhodri, P. Egger 1993, *Some Applications of Recent Developments in Event History Analysis for Historical Demography*, in D.S. Reher and R. Schofield (eds.), *Old and new methods in historical demography*, Oxford, Clarendon Press.

K.A. Doksum, M. Gasko 1990, *On a correspondence between models in binary regression and in survival analysis*, «Int. Stat. Review», 58, 243-2552.

V.T. Farewell 1982, *The use of mixture models for the analysis of survival data with long-term survivors*, «Biometrics», 38, 4, 1041-46.

J. Hajnal 1953, *Age at Marriage and Proportions Marrying*, «Population Studies», 7, 2, 111-36.

A. Hinde 1998, *Demographic Methods*, Arnold Publishers, London.

J. Huang, J. Wellner 1996, *Interval Censored Survival Data: A Review of Recent Progress*, in D. Lin and T. Fleming, (eds.), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, New York, Springer, 123-170.

N.P. Jewell, S.G. Shiboski 1990, *Statistical analysis of HIV infectivity based on partner studies,* «Biometrics», 46, 4, 1133-1150.

N.P. Jewell, M.J. van der Laan 2002, *Current Status Data: Review, Recent Developments and Open Problems*, U.C. Berkeley Division of Biostatistics Working Paper Series, WP 113.

J.P. Klein, M.L. Moeschberger 1997, *Survival Analysis: Techniques for Censored and Truncated Data*, New York, Springer.

L. Li, M.K. Choe 1997, *A mixture model for duration data: analysis of second births in China*, «Demography», 24, 2, 189-197.

R. Maller, X. Zhou 1996, *Survival analysis with long-term survivors*, Chichester, Wiley.

U. Mansmann 2000, *Bayesian methods for the analysis of complex interval-censored event data*, Postdoctoral thesis, Medical School of the Free University of Berlin.

J. McDonald, A. Rosina 2001, *Mixture Modelling of Recurrent Event Times with Long-Term Survivors: Analysis of Hutterite Birth Intervals*, «Statistical Methods & Applications», 10, 257-272.

J.W. McDonald, A.T. Prevost 1997, The fitting of parameter-constrained demographic models, «Mathematical and Computer Modelling», 26, 79-88.

A. Rosina 2000, *Lo studio della fecondità con tecniche di event history analysis. Metodologia ed applicazione ai dati della Seconda Indagine sulla Fecondità in Italia*, CLEUP, Padova.

A. Rosina 2006, A model with long-term survivors for the analysis of current-status nuptiality data, «Population Studies», Vol. 60, No. 1, pp. 73-81.

F. Rossi 1992, L'uso del metodo dei figli propri in demografia storica, «Bollettino di Demografia Storica», 17, 47-70.

K. Yamaguchi 1991, *Event History Analysis*, Sage Pubblications, Newbury Park.

K. Yamaguchi 1992, *Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of 'Permanent Employment' in Japan*, «Journal of the American Statistical Association», 87, 284-92.

K. Yamaguchi, L. Ferguson 1995, *The Stopping and Spacing of Childbirths and Their Birth-History Predictors: Rational-Choice Theory and Event-History Analysis*, «American Sociological Review», 60, 272-298.