

# Università degli Studi di Bologna

---

Dipartimento di Scienze Statistiche  
“P. Fortunati”

## **Gli effetti degli episodi di disoccupazione sulla durata della ricerca di lavoro. Uno studio sui dati del panel europeo**

Tesi di Dottorato di  
Angelina Mazzocchetti

Relatore  
Chiar.mo Prof. Ignazio Drudi

Coordinatore  
Chiar.mo Prof. Daniela Cocchi

Dottorato di Ricerca in  
Metodologia Statistica per la Ricerca Scientifica  
XVI Ciclo



Gli effetti degli episodi di disoccupazione sulla  
durata della ricerca di lavoro.

Uno studio sui dati del panel europeo

Angelina Mazzocchetti

Dipartimento di Scienze Statistiche P. Fortunati

Università degli Studi di Bologna



# Indice

<b>1</b>	<b>La disoccupazione giovanile in Italia: considerazioni generali e scelta dell'ottica di analisi</b>	<b>9</b>
1.1	Il fenomeno analizzato . . . . .	9
1.2	Alcuni aspetti della disoccupazione giovanile in italia . . . . .	11
1.3	La durata della disoccupazione: alcune considerazioni . . . . .	13
<b>2</b>	<b>I dati di durata e i modelli di transizione: rappresentazione e modellizzazione</b>	<b>15</b>
2.1	I dati di durata . . . . .	15
2.1.1	Funzioni caratteristiche di una variabile di durata . . . . .	16
2.2	I modelli di transizione . . . . .	19
2.2.1	Alcune tipologie di modelli condizionali . . . . .	20
2.3	Rappresentazione matematica di un processo puntuale . . . . .	21
2.3.1	Rappresentazione di uno spazio degli stati discreto attraverso un processo di conteggio . . . . .	24
2.3.2	Caratterizzazione della storia di un processo puntuale . . . . .	27
2.3.3	Alcune considerazioni riguardo processi piú generali . . . . .	29
2.4	Definizione e scelta di un modello parametrico . . . . .	31
2.4.1	Test grafici . . . . .	33
2.4.2	L'analisi degli pseudoresidui . . . . .	35
2.5	L'eterogeneità non osservata . . . . .	36
2.5.1	Modelli con mistura di tipo Gamma . . . . .	38
<b>3</b>	<b>Dati utilizzati e loro organizzazione</b>	<b>41</b>
3.1	L'European Community Household Panel (ECHP) . . . . .	41
3.2	Organizzazione dei dati . . . . .	43
3.3	Costruzione delle covariate . . . . .	45
3.3.1	Costruzione delle variabili tempo-dipendenti . . . . .	46
3.4	Caratteristiche personali . . . . .	48
3.5	Caratteristiche famigliari . . . . .	49
3.6	Contesto sociale . . . . .	50

3.7	Il passato del processo . . . . .	50
<b>4</b>	<b>Analisi empirica e risultati</b>	<b>55</b>
4.1	Descrizione degli episodi . . . . .	55
4.2	Scelta della forma parametrica . . . . .	60
4.3	Definizione e stima del modello . . . . .	65
<b>5</b>	<b>Conclusioni</b>	<b>71</b>
<b>A</b>	<b>Struttura del Personal file</b>	<b>77</b>
<b>B</b>	<b>Codifiche per la sezione Calendar of activities</b>	<b>87</b>
<b>C</b>	<b>Matrice di transizione</b>	<b>89</b>
<b>D</b>	<b>Curve di sopravvivenza condizionate.</b>	<b>91</b>
<b>E</b>	<b>Risultati della stima dei modelli</b>	<b>95</b>

# Introduzione

Negli ultimi 20 anni in Europa si è verificato un considerevole aumento del tasso di disoccupazione accompagnato da un aumento ancora più importante per gli individui nella classe di età dai 15 ai 24 anni cioè nei giovani. Nel 1997 il tasso di disoccupazione per questa classe di età era in molti paesi europei più che doppio rispetto a quello delle età adulte [Eurostat 1997, OECD 1998, OECD 1999]. In Italia il tasso di disoccupazione per gli individui sotto i 25 anni era, nel 1997, pari a 33.2% che rappresentava il 12.9 % della popolazione italiana di età 15-25 anni.

In termini assoluti, stando ai dati Eurostat 1998, i giovani, d'età compresa tra i 15 e i 29 anni alla ricerca di una occupazione, sono 1.679.000, ovvero 11mila in più rispetto all'anno precedente, confermando così un andamento in direzione dell'aumento di giovani disoccupati che non sembra aver avuto inversioni di tendenza almeno a partire dal 1992. Il tasso di disoccupazione dei giovani tra i 15 ed i 29 anni conferma quanto anche i dati assoluti ci mostrano: la crescente contrazione delle forze di lavoro in questa fascia d'età, passate da un 7.195.000 giovani nel 1992 a 6.533.000 giovani nel 1996 [Eurostat, 1992, 1996] e l'aumento in termini assoluti dei giovani alla ricerca di un lavoro, conduce, nel corso degli anni '90, all'aumento sistematico del tasso di disoccupazione giovanile.

Questo aumento va guardato come il riflesso di un cambiamento della composizione sociale della disoccupazione verificatosi a partire dagli anni novanta e ancora in corso. In questo periodo infatti si osserva una progressiva contrazione della percentuale dei giovani in cerca di occupazione poiché si verifica uno spostamento verso la disoccupazione in senso stretto (perdita di lavoro) e lo scivolamento dei giovani ancora in attesa di trovare lavoro nelle fasce di età adulte ovvero oltre i 29 anni. Queste considerazioni, accompagnate dalle riflessioni presenti nel par. 1.1 hanno portato alla scelta dei limiti di età 16-34 anni utilizzati per l'analisi empirica.

La realtà sociale, soprattutto in un paese con forti differenziazioni territoriali, generazionali e di genere come l'Italia, è fatta di aspettative, di bisogni, di motivazioni, rispetto l'occupazione, assai diversificate, che tendono a rendere sempre più labile il confine tra lavoro e non lavoro.

Molti sono gli aspetti che caratterizzano, o meglio, che si possono utilizzare per caratterizzare il fenomeno della disoccupazione giovanile. Un aspetto a mio avviso importante è quello della durata della disoccupazione ovvero del periodo di tempo che si impiega nella ricerca di un lavoro.

A partire dai primi anni novanta si è assistiti ad un aumento della percentuale di giovani che cercano lavoro da più di 12 mesi rispetto al numero complessivo dei disoccupati passando dal 56,8 % del 1992 al 67 % nel 1997; l'aumento

si è concentrato soprattutto nella classe di età dai 25 ai 29 anni.

Non va quindi sottovalutata la dimensione di rischio delle giovani generazioni che la disoccupazione può implicare: i giovani che trascorrono gli anni senza trovare un regolare lavoro, quanto più passano dall'adolescenza all'età matura, tanto più subiscono un processo di emarginazione, dagli esiti diversi: esiti che, al limite, sono molto gravi per la loro personalità e per la loro partecipazione sociale.

Data l'importanza della durata della ricerca di lavoro che si accompagna spesso ad esperienze di precarietà si è deciso di analizzare la dipendenza del tasso di passaggio al lavoro dalla durata della ricerca del lavoro stesso ma anche l'effetto che eventuali esperienze vissute nel passato possono avere sulla ricerca attuale di lavoro. L'idea è quindi quella di verificare se sui dati del Panel Europeo sulle Famiglie vi fosse evidenza empirica sull'effetto degli episodi di disoccupazione/occupazione già sperimentati nel passato sulla probabilità " futura " di trovare lavoro.

Gli strumenti metodologici per questo tipo di analisi sono offerti dalla teoria dei dati di durata e dei modelli di transizione già ampiamente utilizzati in letteratura per lo studio della durata della disoccupazione e per la comprensione delle sue possibili determinanti da ricercare nelle caratteristiche personali ma anche familiari e relative al contesto sociale nel quale il fenomeno si esplica. La dinamicità insita in tutti i fenomeni sociali rivestirà un ruolo importante in questo studio: essa verrà esplicitamente considerata sia attraverso variabili esplicative tempo dipendenti sia riconoscendo un ruolo all'evoluzione che la disoccupazione ha avuto nel passato individuale nella convinzione che ciò che siamo nel presente è anche il risultato di ciò che siamo stati nel passato.

Per esplicitare l'effetto dell'esperienza passata rispetto al processo di disoccupazione è stata utilizzata la teoria dei processi stocastici in grado di evidenziare le ipotesi che si fanno nel momento in cui si costruisce il modello di transizione, relative all'effetto del passato sul tasso di transizione. L'idea sviluppata in questa tesi è che l'effetto del passato non sia guidato da ipotesi markoviane o semi-markoviane ma sia riassumibile attraverso una funzione sufficiente per la predizione ovvero che memorizzi in senso dinamico non tutte le informazioni che costituiscono il reale passato dell'individuo rispetto al processo di disoccupazione ma neanche una parte troppo ristretta di essa bensì le informazioni effettivamente importanti. L'importanza degli aspetti considerati è stata determinata su considerazioni generali sul fenomeno ma anche sulla reale disponibilità dei dati utilizzati.

Verranno affrontati essenzialmente due aspetti, quello relativo alla dipendenza del tasso di transizione dalla durata e quello relativo alla stima dei parametri delle variabili che, inserite in un modello di transizione, determinano una variazione più o meno importante dei livelli di tale tasso.



La forma della dipendenza temporale del tasso dalla durata è stata riconosciuta, attraverso test grafici e basati sugli pseudo-residui, essere di tipo log-normale; le variabili esplicative sono state legate al parametro di posizione di tale distribuzione con l'intento di evidenziare le modificazioni che alla durata media dell'episodio di disoccupazione vengono apportate dalle variabili esplicative. Tra queste variabili esplicative verrà inserita la storia del processo di disoccupazione. Dopo alcune notizie relative alla disoccupazione giovanile in Italia negli anni novanta, nel capitolo 2 viene presentato l'assetto metodologico utilizzato in fase di stima. Segue una breve descrizione dei dati utilizzati e delle variabili costruite per finire, nel capitolo 4 con la presentazione dei risultati dell'applicazione empirica.



# Capitolo 1

## La disoccupazione giovanile in Italia: considerazioni generali e scelta dell'ottica di analisi

### 1.1 Il fenomeno analizzato

In questo studio si intende analizzare alcuni aspetti legati al fenomeno della disoccupazione giovanile in Italia.

Il fenomeno della disoccupazione e specialmente quello della disoccupazione giovanile influenza la struttura del mercato del lavoro ma ha anche importanti riflessi nella società nel suo complesso; nella nostra società, che si basa per la maggior parte su un sistema di reddito proveniente dal lavoro dipendente, essere esclusi dal mercato del lavoro può avere seri riflessi su vari aspetti della vita degli individui. Il lavorare è correlato con l'integrazione e l'esistenza sociale: la disoccupazione, e in particolar modo quella di lunga durata, causa problemi economici e può anche determinare la deteriorazione delle attrattive dell'individuo nel mercato del lavoro rappresentando quindi il primo passo verso l'esclusione sociale e la povertà.

Il risultato di questo processo è la creazione dei cosiddetti lavoratori scoraggiati: la lunga infruttuosa attesa durante la ricerca di lavoro sembra rappresentare l'impossibilità di trovarlo e questo produce scoraggiamento riguardo il mercato del lavoro.

La possibilità di studiare un fenomeno passa attraverso una definizione il più precisa e univoca possibile del fenomeno stesso, del quale, in ogni caso, difficilmente comprenderemo le dinamiche di fondo se non relativamente ad aspetti ristretti e ben identificati dello stesso.

In questo studio risulta innanzitutto fondamentale la definizione di *disoccu-*

*pato* che permette l'identificazione dei soggetti da analizzare. Una definizione molto utilizzata nelle indagini di tipo longitudinale è quella data dall'Organizzazione Internazionale del Lavoro (International Labour Organization ILO). Un individuo è disoccupato se non ha lavorato più di un'ora in un breve periodo di riferimento (generalmente la settimana o il giorno precedente quello dell'intervista), che sia immediatamente disponibile a lavorare e che stia attivamente cercando lavoro.

Definita la condizione di disoccupato dobbiamo identificare la classe di riferimento per i *giovani*: lo standard è di considerare giovani le persone di età tra i 15 e i 24 anni inclusi ma nella realtà di ogni paese differenti fattori, istituzionali culturali o politici possono influenzare questi limiti.

Nello studio che verrà presentato il campione di riferimento è costituito dai partecipanti all'European Community Household Panel (ECHP) di età tra i 17 e i 34 anni. I limiti sono stati definiti da un lato, analizzando la situazione socio-demografica e rispetto all'ingresso nel Mercato del Lavoro dei giovani italiani, e dall'altro, attraverso l'evidenza empirica riscontrata sui dati.

Se consideriamo che in Italia il periodo che si trascorre mediamente in istruzione, per i livelli superiori, è circa 7 anni e che si possono avere dei picchi di 10 anni per alcune facoltà come ingegneria o medicina sembra già ragionevole considerare il limite superiore attorno ai 30-35 anni.

Alcune evidenze empiriche [Billari, Ongaro, 1999] ci dicono che intorno ai 30-35 anni si verifica la prima completa indipendenza dalla famiglia d'origine e che questa è spesso collegata alla possibilità di trovare un lavoro stabile. Sappiamo anche che tra il 1986 e il 1994, la percentuale degli uomini italiani tra i 25-29 che vivono nella casa dei genitori è aumentata considerevolmente passando dal 49.6% al 66%. Il livello corrispondente per le donne è inferiore, risultato della preferenza delle donne a sposare uomini più grandi. Ciò nonostante esse hanno sperimentato parallelamente agli uomini un incremento dal 25.5% al 44.1% [Cordon, 1997]. L'autore suggerisce, per queste evidenze, una spiegazione economica basata sul drammatico aumento nello stesso periodo del tasso di disoccupazione generale nonché giovanile e sull'elevato costo delle abitazioni sia in vendita che in affitto.

Sulla scelta dei limiti di età da considerare ha giocato in ogni caso anche l'evidenza empirica. La classe di età 16-25/29 anni individuava la presenza di molti giovani ancora coinvolti nel processo di istruzione e che facevano registrare degli episodi di occupazione saltuaria, ad esempio stagionale, e inserire comunque nel processo di istruzione; questi individui non sono disoccupati in quanto non ancora formalmente inseriti nel Mercato del Lavoro. Ne segue anche che gli individui disoccupati in questa fascia di età hanno principalmente titoli di istruzione inferiori e si muovono come precari all'interno del Mercato del Lavoro.

## 1.2 Alcuni aspetti della disoccupazione giovanile in Italia

Molti fattori possono essere responsabili della durata e del numero di episodi nel corso di vita un giovane.

I dati utilizzati nell'analisi qui presentata riguardano il periodo dal 1994 al 1998, cerchiamo quindi di rilevare quali elementi hanno caratterizzato la disoccupazione giovanile in Italia attorno a quegli anni.

In generale nei paesi sviluppati, e quindi anche in Italia, si è osservato nel corso degli ultimi 15 anni un effetto forte e contrastante del livello di istruzione sulla probabilità di trovare lavoro e sulla durata complessiva della ricerca stessa.

I giovani, anche se con maggior educazione che in passato, sono sempre più spesso disoccupati e generalmente sperimentano un periodo di forte instabilità nel mercato del lavoro prima di essere assunti con un contratto stabile. La posizione dei giovani rispetto al sistema scolastico e al mercato del lavoro varia considerevolmente tra gli stati europei ma si è osservato in tutti un aumento dei livelli di istruzione raggiunti con una diminuzione del tasso di partecipazione al mercato del lavoro; post-porre l'ingresso nel mercato del lavoro può essere visto come un modo per acquisire maggiore qualificazione e migliorare le prospettive in ambito lavorativo. Contemporaneamente si osserva un incremento nella difficoltà di entrata nel mercato del lavoro per gli individui meno qualificati dovuta all'alta competizione per i (pochi) posti disponibili e quindi un generale peggioramento delle condizioni di questo gruppo di persone.

Dalle analisi condotte sembra che la transizione al lavoro dipenda fortemente dal livello di educazione raggiunto: finire almeno l'istruzione secondaria superiore sembra essere una condizione necessaria, ma non sufficiente, per trovare un lavoro. In alcuni paesi come l'Italia, la Spagna, la Grecia e il Portogallo molti giovani tra i più educati devono aspettare un tempo maggiore per trovare un lavoro [OECD,1999] anche se in generale a restare nella condizione di disoccupato per periodi più lunghi sono quelli con livelli d'istruzione inferiori.

Tra il 1993 e il 1999 il sistema economico ha creato nuovi posti di lavoro per oltre un milione di unità nelle professioni non manuali, quasi esclusivamente in quelle ad alta qualifica, e ne ha distrutti quasi altrettanti nelle professioni manuali. Dal rapporto annuale Istat [Istat,1999] sulla situazione del paese nel 1999 leggiamo che tra l'ottobre del 1992 e il gennaio del 2000 l'incidenza del lavoro atipico, intendendo con questo termine tutte le forme contrattuali a tempo determinato che non danno nessuna sicurezza rispetto all'evoluzione

futura dell'occupazione stessa, è passata dal 10,6% al 15,2 % e che fino al 1997 la diffusione di queste forme di lavoro è avvenuta a scapito dell'occupazione standard, almeno a livello aggregato ovvero non considerando i possibili sotto-settori del lavoro nei quali l'incidenza potrebbe essere stata differente.

Se consideriamo i sotto-settori di attività economica vediamo che il lavoro atipico ha caratterizzato maggiormente il commercio, l'agricoltura e i servizi di mercato e che ha riguardato principalmente le donne e i giovani.

Dall'andamento di flusso del problema della disoccupazione generale appare subito evidente il peso del dualismo territoriale, ed il prevalente carattere femminile del fenomeno.

Insieme alla marcata distanza che separa il dato della disoccupazione nel Centro-Nord, inferiore alla stessa media della Comunità Europea, dal dato della disoccupazione del Sud, che ammonta a circa tre volte quello relativo alle regioni centro-settentrionali, in prospettiva storica, abbiamo che la disoccupazione al Sud, con poche eccezioni, aumenta anche quando al Nord ed al Centro si assiste ad un suo arresto o arretramento.

Disaggregando ulteriormente i dati su base territoriale, infatti, secondo i dati Istat relativi al 1997, le regioni meridionali accusano un aumento della disoccupazione rispetto ai primi anni novanta, passando dal 21.7 al 22.2 % mentre il tasso diminuisce leggermente per quanto concerne il Centro (dal 10.3 al 10.2 % ) e rimane invariato al Nord (6.6 % ).

Secondo i dati Istat (1997) il principale contributo all'aumento della disoccupazione complessiva nel 1996 venuto dai giovani in cerca di prima occupazione, aumentati del 4.7 % aumento più sostenuto al Sud che al Nord.

Di fatto la componente giovanile della disoccupazione diminuisce relativamente al dato complessivo della disoccupazione, della quale costituiva il 69.1 % nel 1992 ed il 60,7 % nel 1996, a testimonianza di una dinamica generale che incide sulla composizione sociale della disoccupazione e che si riflette anche su quella giovanile: la crescita del peso relativo della disoccupazione in senso stretto (coloro che hanno perso una precedente occupazione), e lo scivolamento dei giovani senza lavoro nella fascia d'età degli adulti, cioè oltre i 29 anni, ancora in attesa di trovare una occupazione. Va, dunque, rilevata la nuova qualità della disoccupazione giovanile, che negli anni '90 assume crescente importanza come esclusione dal mondo del lavoro e riduce invece il peso della componente dei giovani in cerca di prima occupazione. Ma anche questo dato va scomposto tenendo conto del dualismo territoriale come variabile decisiva per la comprensione del fenomeno analizzato. In un quadro nazionale complessivo comunque caratterizzato dalla ridotta presenza di disoccupati in senso stretto a fronte di una schiacciante prevalenza di persone in cerca di primo lavoro, mentre nel Centro-Nord assistiamo, a partire dalla metà degli

anni '80, ad un mutamento nel mercato del lavoro, per cui è più facile per i giovani trovare impiego, che però spesso risulta precario, quindi non pochi diventano disoccupati in senso stretto, nel Sud, invece, questo mutamento si verifica in modo più contenuto. Anzi, in quest'area del paese, le difficoltà per un primo inserimento lavorativo paiono intensificarsi, tanto che si assiste ad un forte invecchiamento delle persone in cerca di prima occupazione e ad una crescente dilatazione della fascia di giovani meridionali che entra nella vita adulta senza aver mai avuto un'esperienza lavorativa significativa, che non sia fatta cioè di lavoretti mal pagati, precari o irregolari.

### **1.3 La durata della disoccupazione: alcune considerazioni**

Un aspetto per me particolarmente interessante da analizzare è quello della durata della disoccupazione ovvero del tempo che un individuo passa in cerca di lavoro. Abbiamo visto come molto tempo venga dedicato alla ricerca del primo lavoro e anche come questo si sia spostato sempre di più verso le fasce di età adulte, dopo i 29 anni tra i quali troviamo però anche sempre più disoccupati in senso stretto ovvero che hanno perso il lavoro a causa della crescente precarizzazione del lavoro.

I giovani alla ricerca di prima occupazione costituiscono la componente più rilevante della disoccupazione giovanile, e più precisamente il 58 % di essa [Istat,1997]. La probabilità di questi giovani alla ricerca della prima occupazione di trovare lavoro nell'arco di un anno è pari al 20 % la metà di quella del complesso dei disoccupati. Questo spiega anche la prevalenza, tra i giovani disoccupati, di coloro che cercano un lavoro da più di 12 mesi: essi costituiscono il 67 % dei disoccupati in tale fascia d'età, con un aumento di dieci punti in percentuale rispetto al 1992.

La crescita della percentuale di giovani che cercano lavoro da più di 12 mesi ha riguardato soprattutto i soggetti tra i 25 ed i 29 anni: mentre il peso dei 15-24enni disoccupati di lunga durata, relativamente al numero complessivo dei giovani disoccupati, aumenta di un punto in percentuale tra il 1992 ed il 1997, quella dei 25-29enni aumenta di ben 9 punti percentuali sia a causa della sempre minore capacità del mercato del lavoro di assorbire nuovi lavoratori sia per il posticipo dell'ingresso vero e proprio dei giovani in esso. Se da un lato giocano molti fattori determinati dai meccanismi interni al mercato del lavoro dobbiamo anche considerare che il tempo d'attesa per gli individui con alti livelli d'istruzione è in realtà dipendente solo in parte da tali meccanismi: per questi individui infatti si ipotizza anche un meccanismo

di auto-rinuncia ad alcuni lavori guidato dalle aspettative rispetto a quello che è un *buon lavoro* che si riflette in una attesa volontaria per ricercare il lavoro maggiormente correlato agli studi fatti.

Negli ultimi anni, alla diminuzione del tasso globale di disoccupazione nell'UE si è accompagnata una lenta flessione del numero dei disoccupati di lunga durata, ma questo fenomeno ha riguardato l'Italia solo in maniera marginale. Nel 1996 il tasso di disoccupazione di lunga durata non era di molto inferiore a quello del 1994 ed il numero dei disoccupati, era inferiore solo di qualche punto percentuale rispetto ai 4 anni precedenti in cui la disoccupazione era al suo picco. La percentuale di disoccupati da almeno un anno o più (59,5 % ) era superiore a quella del 1994 (47 % ) . Tali valori indicano che i disoccupati di lunga durata devono affrontare enormi difficoltà per trovare un posto di lavoro anche quando la disoccupazione è in calo. Il numero di disoccupati di lunghissima durata , ovvero coloro che sono alla ricerca di un posto di lavoro da 2 o più anni, era infatti superiore nel 1998 rispetto al 1994 (con aumento superiore al 5 % ) e, il loro peso relativo rispetto all'intera massa dei disoccupati, di gran lunga superiore.

Bisognerebbe, dunque, non sottovalutare la dimensione di rischio delle giovani generazioni che la disoccupazione può implicare: i giovani che trascorrono gli anni senza trovare un regolare lavoro, quanto più passano dall'adolescenza all'età matura, tanto più subiscono un processo di emarginazione, dagli esiti diversi: esiti che, al limite, sono molto gravi per la loro personalità e per la loro partecipazione sociale.



## Capitolo 2

# I dati di durata e i modelli di transizione: rappresentazione e modellizzazione

### 2.1 I dati di durata

Per dato di durata si intende qualsiasi dato che raccolga l'informazione sulla durata di un determinato evento ovvero il tempo che una unità statistica, sia essa un individuo o una azienda, trascorre in un determinato stato.

L'interesse sarà quindi rivolto all'analisi della legge distributiva di una variabile aleatoria non negativa. Per ottenere una definizione non ambigua della durata è necessario definire univocamente il momento di origine dell'asse temporale ( $t = 0$ ) e la scala con la quale tale tempo viene misurato.

Se consideriamo un individuo possiamo definire e analizzare il tempo trascorso da disoccupato prima di trovare un lavoro e interrogarci su quali sono i fattori che in qualche modo hanno influenzato questo cambiamento di stato oppure, più semplicemente, come nei modelli marginali, analizzare come la probabilità di cambiare stato dipende dal tempo trascorso in esso.

Se osserviamo l'individuo per un periodo di tempo abbastanza lungo si possono avere più durate di disoccupazione intervallate da periodi trascorsi in altri stati dello stesso processo, ad esempio, lavorando oppure seguendo un corso di formazione. Affiancando a queste durate quelle che l'individuo trascorre in stati che descrivono altri processi, ad esempio quello relativo allo studio oppure alla vita coniugale è possibile, in teoria, avere una descrizione dell'intera vita di un individuo attraverso la sequenza degli stati attraversati e quella delle durate trascorse in essi.

Se si ha la possibilità di osservare l'intera durata di vita di un individuo è

possibile avere dati in forma completa ovvero associare ad ogni stato una durata completamente osservata di permanenza in quello stato; in molti casi, come con i dati utilizzati per l'applicazione qui presentata, non si conosce l'intero corso di vita ma solo parte di questo e quindi le durate osservate possono non essere complete. In questi casi si dovranno trattare durate dette censurate a sinistra oppure a destra.

Gli episodi censurati a destra sono quelli per cui non si conosce il tempo finale perchè successivo all'ultimo istante di osservazione. Se la transizione avviene é in un momento esterno alla finestra di osservazione. per le censure a sinistra quello che non si conosce é il momento iniziale dell'episodio ma solo quello finale: l'osservazione inizia quando l'individuo occupa già lo stato al quale siamo interessati e non sappiamo quando ci è entrato.

In entrambi i casi le durate osservate non sono complete e quindi il contributo che danno alla conoscenza del fenomeno è incompleto e in fase di analisi sarà necessario fare delle ipotesi sui meccanismi che le hanno generate.

Per analizzare, nel senso di comprendere, i meccanismi che determinano la probabilità di passare da uno stato ad un altro, se e in che modo questa dipenda dalla durata trascorsa in tale stato <sup>1</sup>, quali caratteristiche la influenzano, é necessario avere una rappresentazione delle informazioni a disposizione che sia utile e maneggevole da un punto di vista matematico e statistico e non solo fenomenico.

Lo strumento che permette questo tipo di riflessione é il modello di transizione. Rappresentando il processo osservato come un processo stocastico si arriverá ad una rappresentazione di tali modelli in grado di rendere visibili le ipotesi rispetto all'influenza del passato stesso del processo.

### 2.1.1 Funzioni caratteristiche di una variabile di durata

Per descrivere la legge di probabilità di una variabile aleatoria facciamo ricorso alla sua funzione di distribuzione  $F_T(t)$  che esprime la probabilità che la durata osservata sia inferiore o uguale ad un certo valore  $t$ ;  $F_T(t) = P(T \leq t)$  è la probabilità che l'evento di interesse avvenga prima dell'istante  $t$ . Definendo la funzione di distribuzione é possibile, attraverso la sua derivata rispetto al tempo, definire anche la funzione di densità di probabilità  $f_T(t)$ .

La funzione che realmente caratterizza i dati di durata é la funzione di sopravvivenza  $S_T(t)$ . Essa esprime la probabilità di sopravvivere all'evento

---

<sup>1</sup>Ci si riferisce alla dipendenza temporale del tasso ovvero al fatto che il tasso non é considerato costante nel tempo ma variabile a seconda del valore della durata

fino al tempo  $t-$  cioè la probabilità che la durata osservata sia maggiore di  $t$ :  $S_T(t) = P(T \geq t)$ . Questa funzione, nel caso assolutamente continuo, rappresenta esattamente il complemento della funzione di distribuzione.<sup>2</sup> Le proprietà caratteristiche di questa funzione sono:

- i )  $S_T(t) \in [0, 1]$       $t \in \mathbb{R}^+$
- ii )  $S_T(t)$  é continua a sinistra i.e.  $S_T(t-) = S_T(t)$       $t \in \mathbb{R}^+$
- iii )  $S_T(t)$  é monotona decrescente i.e.  $t_1 > t_2 \implies S_T(t_1) \leq S_T(t_2)$
- iv )  $S_T(0) = 1$  e  $\lim_{t \rightarrow \infty} S_T(t) = 0$

Attraverso la funzione di sopravvivenza si definisce un modello elementare marginale, che non prevede cioè l'esistenza delle covariate e che relaziona la probabilità di accadimento dell'evento solo alla durata osservata. Stimando la funzione di sopravvivenza é possibile ottenere ad esempio la durata mediana e avere quindi indicazioni sulla velocità di eliminazione degli individui del campione ovvero un'idea della dipendenza temporale delle transizioni. In realtà la dipendenza temporale della transizione dalla durata viene evidenziata attraverso la stima della funzione di rischio riconducibile al problema della legge di probabilità di  $T$  condizionatamente a  $T \geq t$ . La funzione di rischio  $h_T(t)$  può essere derivata dalla funzione di sopravvivenza attraverso la seguente relazione:

$$h_T(t) = -d \ln S_T(t) / dt$$

Tra le funzioni definite,  $F_T(t)$ ,  $f_T(t)$ ,  $S_T(t)$ ,  $h_T(t)$  esistono relazioni ben precise che fanno sì che a partire dalla conoscenza di una di esse sia possibile ricavare anche tutte le altre, vedi ad esempio [Kalbfleish, Prentice, 1980], [Lancaster, 1990], [Lawless, 1982].

I modelli marginali possono essere utilizzati in fase descrittiva ad esempio stimando una funzione di sopravvivenza per ogni strato del campione definito dall'incrocio di diverse caratteristiche. Test comparativi di tali funzioni possono indicare se la variabile di stratificazione considerata é o meno informativa nel senso che determina una differenza significativa nel comportamento dei sotto-campioni che definisce.

I test che verranno utilizzati sono due: Log-Rank (Savage) e il test di Wilcoxon in tre diverse formulazioni rispettivamente di Breslow, Tarone-Ware e

---

<sup>2</sup>Ciò é vero solo in tempo continuo dove la probabilità dell'evento  $T = t$  é nulla. In tempo discreto

$$S_T(t) = P(T \geq t) = 1 - F_T(t) + P(T = t)$$

Prentice. Tutti questi test si basano sull'ipotesi nulla che la funzione di sopravvivenza non differisce tra i gruppi considerati ma hanno una sensibilità diversa e a volte danno risultati discordanti. La ragione risiede nel fatto che il test di Wilcoxon (nelle tre formulazioni) esalta le differenze presenti all'inizio dell'asse temporale mentre il Log-Rank test evidenzia le differenze approssimativamente dalla metà alla fine dell'asse temporale (a tale proposito si veda [Blossfeld,Rohwer,1995] figura 3.3.2 pag.75). Questi test hanno la stessa idea di base ma si differenziano per l'utilizzo di differenti sistemi di pesi. Guardiamo brevemente come questi test vengono costruiti: si definiscono  $\mathbf{m}$  gruppi e si calcolano le quantità per riempire la seguente tavola

$\tau_1$	$R_{11}$	$E_{11}$	$R_{12}$	$E_{12}$	$\dots$	$R_{1m}$	$E_{1m}$
$\tau_2$	$R_{21}$	$E_{11}$	$R_{22}$	$E_{22}$	$\dots$	$R_{2m}$	$E_{2m}$
$\tau_3$	$R_{31}$	$E_{11}$	$R_{32}$	$E_{32}$	$\dots$	$R_{3m}$	$E_{3m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\dots$
$\tau_q$	$R_{q1}$	$E_{q1}$	$R_{q2}$	$E_{q2}$	$\dots$	$R_{qm}$	$E_{qm}$

dove  $\tau_i$  con  $i \in \{1, 2, \dots, q\}$  sono i punti nel tempo dove si verifica almeno un evento,  $E_{ij}$  con  $i \in \{1, 2, \dots, q\}$  e  $j \in \{1, 2, \dots, m\}$  è il numero di episodi del gruppo  $j$  che hanno un evento in  $\tau_i$  e  $R_{ij}$  è il risk-set in  $\tau_i$  ovvero l'insieme degli episodi del gruppo  $j$  con tempo iniziale inferiore a  $\tau_i$  e tempo finale maggiore di  $\tau_i$ .

I test statistici  $\mathcal{T}_\nu$  con  $\nu = 1, 2, 3, 4$ , sono definiti come:

$$\mathcal{T}_\nu = U^{(\nu)'} V^{(\nu)-1} U^{(\nu)}$$

dove  $U^{(\nu)}$  è un vettore  $m$ -dimensionale e  $V^{(\nu)}$  è una matrice di dimensioni  $m^*m$  con elementi per ogni gruppo dati da:

$$U_j^{(\nu)} = \sum_{i=1}^q W_i^{(\nu)} (E_{ij} - R_{ij} \frac{E_{i0}}{R_{i0}})$$

e

$$V_{j_1, j_2}^{(\nu)} = \sum_{i=1}^q W_i^{(\nu)^2} \frac{E_{i0}(R_{i0} - E_{i0})}{R_{i0} - 1} \frac{R_{ij_1}}{R_{i0}} \left( \mathbb{I}_{[i=j]} - \frac{R_{ij_2}}{R_{i0}} \right)$$

$W_i^{(\nu)}$  rappresenta il diverso sistemi di pesi definiti per  $i \in \{1, 2, \dots, q\}$  come:

- $W_i^{(1)} = 1$  Log-Rank test
- $W_i^{(2)} = R_i$  Wilcoxon-Breslow

- $W_i^{(3)} = \sqrt{R_i}$  Wilcoxon-Tarone-Ware
- $W_i^{(4)} = \prod_{k=1}^i \frac{R_i - E_i + 1}{R_i + 1}$  Wilcoxon-Prentice

Tutti i test seguono una distribuzione  $\chi^2$  con  $m - 1$  gradi di libertà sotto l'ipotesi nulla che non vi siano differenze significative. Notiamo anche che il rango di  $V^{(\nu)}$  è  $m-1$  quindi il calcolo dei test richiede l'utilizzo dell'inversa generalizzata.

L'operazione di confrontare le curve di sopravvivenza al fine di scoprire quali fattori le modificano significativamente diventa difficile quando ci sono molte variabili e in ogni caso fornisce indicazioni che potrebbero non essere vere nel momento in cui si tiene conto anche della presenza di altri fattori che agiscono contemporaneamente a quello usato per la stratificazione.

Dai modelli marginali di dipendenza pura della transizione dalla durata osservata, l'interesse si sposta a questo punto sui modelli condizionali ovvero preposti alla stima della legge di probabilità della durata condizionatamente ad un set di covariate: variabili endogene che possono influenzare la probabilità di transizione.

Per specificare i modelli condizionali si può considerare un individuo di riferimento e vedere cosa rende gli altri individui diversi da lui. Nei modelli parametrici si sceglie una distribuzione di riferimento per la funzione di rischio di tale individuo <sup>3</sup> e si legano le covariate ai parametri di tale distribuzione; lasciando completamente non specificata la forma di tale funzione si ottengono modelli non parametrici mentre specificando parametricamente solo la forma dell'influenza delle covariate ma non quella distributiva della funzione di rischio si vanno a stimare modelli semi-parametrici.

## 2.2 I modelli di transizione

Un *modello di transizione generale*<sup>4</sup> rappresenta una sequenza di stati e di durate in quegli stati, intendendo con il termine generale che più di una transizione è possibile all'interno dello spazio degli stati. Ne risulta che gli elementi costitutivi un modello di transizione sono uno spazio degli stati  $E = \{E_0, E_1, E_2, \dots, E_p\}$  e una sequenza di durate  $(T_1, T_2, \dots, T_p)$ .

Attraverso tale strumento è possibile modellizzare la probabilità istantanea di entrare nello stato  $E_i$  al tempo  $t$  dato che si era nello stato  $E_j$  fino al tempo

---

<sup>3</sup>Detta anche rischio di base

<sup>4</sup>Anche detto modello di durata in quanto la variabile che si va a modellizzare è una durata che termina o meno con una transizione

$t - 1$  esattamente precedente la transizione. Questa probabilità condizionata é esattamente data dal tasso di transizione che permette di indagare la dipendenza del processo osservato dal tempo, ovvero del momento della transizione dalla durata osservata, e di legarla a delle covariate informative di quelle caratteristiche che, a parit  di durata, rendono la transizione pi  o meno probabile.

La seguente trattazione considerer  un processo con spazio degli stati discreto e tempo continuo avvalendosi della rappresentazione dello stesso processo in tempo discreto.

In questo tipo di modelli l'attenzione si rivolge alla legge di probabilit  che domina la distribuzione della variabile aleatoria non-negativa che rappresenta la durata. Il modello di transizione pi  elementare   rappresentato da una singola durata ovvero da un singolo episodio di interesse che termina con un dato tipo di transizione. La struttura di modelli pi  complessi che prevedono quindi l'esistenza di pi  durate e pi  transizioni per un singolo individuo in un dato intervallo di tempo pu  essere ottenuta per estensione del modello elementare.

### 2.2.1 Alcune tipologie di modelli condizionali

Nei modelli parametrici si sceglie una distribuzione di riferimento per la variabile di durata, i parametri di tale distribuzione verranno stimati direttamente sui dati e determineranno esattamente la forma del tasso di transizione.

In teoria ogni distribuzione non-negativa pu  essere utilizzata per parametrizzare dei dati di durata anche se, in pratica, le distribuzioni pi  utilizzate sono le seguenti :esponenziale, Gamma, Weibull, Gompertz-Makeham, log-normale e log-logistica. La scelta della forma parametrica   guidata da considerazioni di carattere diverso provenienti sia da riflessioni teoriche sia da precedenti conoscenze del fenomeno sia dal confronto attraverso appositi test dell'adattamento di funzioni di rischio parametriche alla stima della stessa con metodi non parametrici come quello di Kaplan-Meyer.

Nei modelli non parametrici si usano tali metodi per ottenere una stima della distribuzione delle durate dalla quale, attraverso opportune trasformazioni, ottenere una stima della funzione di rischio.

I modelli semi-parametrici rappresentano il ponte d'unione tra quelli parametrici e non in quanto lasciano non specificata la forma del rischio di base e quindi non fanno nessuna assunzione sulla sua dipendenza temporale ma ne specifica invece una per l'effetto delle covariate. Il modello semi-parametrico pi  noto   senz'altro il modello di Cox che lascia il rischio di base non speci-

ficato e specifica un effetto esponenziale per le covariate.

Una importante ulteriore distinzione é definita attraverso il trattamento che si fa dell'effetto di cambiamento indotto da una covariata. In linea di massima si distinguono due grandi classi: i modelli a rischio proporzionale e quelli a tempo accelerato.

In generale nel primo gruppo rientrano quei modelli per cui si definisce un rischio di base  $h_0(t|\alpha)$ <sup>5</sup> proprio dell'individuo senza nessun trattamento e la funzione di rischio  $h_T(t|\theta)$ <sup>6</sup> specifica che l'effetto di un regressore é di moltiplicare questo rischio di base per un coefficiente costante non noto  $\beta$  che incrementa o decrementa il rischio di base. Il risultato é che le funzioni di rischio di due individui che differiscono per l'avere o non avere una determinata caratteristica sono parallele e hanno un coefficiente di proporzionalitá  $\beta$  costante per tutte le durate: l'effetto delle covariate é di moltiplicare il tasso di base per una costante che farà aumentare ( $\beta > 0$ ) o diminuire ( $\beta < 0$ ) il rischio di base:  $h_T(t|\theta) = \beta h_0(t|\alpha)$ .

Nel secondo gruppo di modelli l'effetto delle covariate accelera o decelera l'accadimento dell'evento. La trasformazione effettuata puó essere vista come un cambiamento di scala della variabile tempo; l'effetto dei regressori é di cambiare scala alla variabile di durata nel senso di evidenziare l'anticipo o il ritardo nel vivere l'evento per gli individui che hanno o meno una determinata caratteristica.

Per la definizione di un modello statistico è importate esplicitare in maniera univoca la sua struttura ma anche quella dei dati ai quali il modello si applica: ciò permette sia l'identificazione delle sue proprietà sia la possibilità di interpretare i risultati.

Nei paragrafi successivi vedremo come la teoria dei processi stocastici fornisca uno strumento molto utile per la definizione dei modelli di transizione e delle sue proprietà nonché del modo in cui l'evoluzione passata del processo in studio, ovvero la storia individuale rispetto al fenomeno " disoccupazione " entra nel modello e che ruolo riveste.

## 2.3 Rappresentazione matematica di un processo puntuale

Consideriamo  $X_t$  un processo stocastico in tempo continuo con stato degli stati discreto:

$$X = \{X_t : t \geq 0\} \quad X_t \in E = \{E_0, E_1, E_2, \dots, E_p\}$$

---

<sup>5</sup> $\alpha$  rappresenta l'eventuale parametro in una formulazione parametrica

<sup>6</sup> $\theta = (\alpha, \beta)$  dove  $\beta = (\beta_1, \beta_2, \dots, \beta_r)$  e  $r$  é il numero dei regressori

dove  $E$  rappresenta lo spazio degli stati, insieme di tutti gli stati possibili per quel processo, e  $t$  appartiene ad un intervallo non-degenerato della retta reale positiva:

$$t \in I \quad \text{dove} \quad I = [0, a] \quad \text{oppure} \quad I = R_+ = [0, \infty)$$

Con questa struttura possiamo definire una traiettoria,  $x$ , come una funzione misurabile definita in  $I$  con valori,  $x_t$ , nello spazio degli stati  $E$ :

$$x : I \rightarrow E : \forall t \in I \mapsto x_t \in E$$

ovvero  $x$  appartiene allo spazio funzionale  $E^I$  di tutte le possibili funzioni da  $I$ , asse temporale, ad  $E$ , spazio degli stati;  $x_t$  ci dice in quale stato si trova l'individuo per ogni istante di tempo.

Possiamo ora definire un processo puntuale  $X$  come una funzione definita in uno spazio probabilistico che prende valori in uno spazio di traiettorie. Dato lo spazio probabilistico  $(\Omega, A, P)$  definiamo:

$$X : \Omega \rightarrow E^I \quad \forall \omega \in \Omega \mapsto X(\omega) = (X_t(\omega) : t \in I) \in E^I$$

cióe per ogni  $t$  abbiamo una variabile aleatoria che crea una famiglia di variabili aleatorie indicizzate da  $t \in I$ .

Per avere una buona struttura è necessario richiedere che le traiettorie  $X(\omega)$  siano misurabili ovvero che la loro legge di probabilità sia derivabile da  $P$  : rendere un processo puntuale operativo non é necessario ad evitare situazioni nelle quali, ad esempio, il numero di transizioni sia non numerabile <sup>7</sup>.

Una condizione di regolarità sufficiente [Florens et al., 1994] per la misurabilità di questa funzione è che le traiettorie siano, con probabilità uno funzioni **CADLAG** del tempo: continue a destra e con limite sinistro esistente. In termini formali diciamo che una funzione ha questa proprietà quando:

$$X(t_+) = X(t) \quad \text{e} \quad X(t_-) \text{ esiste.}$$

Questa proprietà assicura che:

- il numero delle transizioni è numerabile e forma un insieme ordinato; possiamo usare l'insieme dei numeri naturali per indicizzare il numero delle transizioni
- Le funzione di  $X$  ed anche quelle degli istanti di transizione e delle durata trascorsa in quello stato sono misurabili quindi la loro legge di probabilità ben definita: si può passare da  $\Omega$  a  $E^I$ .

---

<sup>7</sup>per un esempio di processo con un numero di transizioni non numerabile ma legge di probabilità perfettamente definita si veda [Mouchart,2003]



- Gli istanti nei quali si verificano le transizioni sono ordinati nel tempo con probabilità uno.
- Le durate sono strettamente positive con probabilità uno.

In conseguenza di questa proprietà una traiettoria completa definita su un asse temporale continuo ammette una rappresentazione equivalente in tempo discreto senza perdita di informazioni; il processo in tempo continuo  $(X_t : t \in I)$  può essere rappresentato con un processo bivariato in tempo discreto.

A partire dal processo  $X_t$  definiamo i seguenti processi discreti:

- i ) Un processo per il punto nel tempo nel quale si verifica un cambiamento di stato definito ricorsivamente come :

$$\tau_0 = 0, \quad \tau_p = \inf\{t | t > \tau_{p-1}, X_t \neq X_{\tau_{p-1}}\}$$

quindi  $\tau_p : \Omega \rightarrow R_+ \quad \forall p \in N \quad \text{con} \quad \tau_p > \tau_{p-1}$  quasi certamente

- ii ) un processo per le durate indicante da quanto tempo il processo si trova nello stesso stato definito ricorsivamente come:

$$T_0 = 0 \quad T_p = \tau_p - \tau_{p-1} \quad \text{con} \quad T_p > 0 \quad \text{quasi certamente}$$

Nello specifico si identificheranno con  $T_p$  la durata nello stato  $X_{\tau_{p-1}}$  tra le transizioni (p-1) e (p).

Con questa rappresentazione abbiamo che l'informazione data dalla completa osservazione del processo continuo  $X_t$  equivalente a quella data dalla traiettoria completa del processo discreto

$$(X_t : t \in I) \iff \{(\tau_p, X_{\tau_p}) : p \in N\} \quad \text{dove} \quad (\tau_p, X_{\tau_p}) \in R_+ \times E$$

ovvero

$$\left\{ \left( T_p, X \left( \sum_{i=0}^p T_i \right) \right) \right\} \quad \text{dove} \quad \left( T_p, X \left( \sum_{i=0}^p T_i \right) \right) \in R_+ \times E$$

indicando con  $\iff$  l'equivalenza delle informazioni, ovvero dei  $\sigma$ -campi generati dai due processi <sup>8</sup>.

Il passaggio da una rappresentazione nel continuo ad una discreta - da non numerabile a numerabile- rappresenta il primo passo per poter trattare statisticamente un processo continuo con spazio degli stati discreto, ma non sufficiente poiché abbiamo considerato, finora, uno spazio degli stati senza nessuna struttura matematica .

Per raggiungere questo obiettivo si utilizzerà un processo di conteggio per rappresentare lo spazio degli stati discreto.

---

<sup>8</sup>Osserviamo che  $\tau_p = \sum_{i=0}^p T_i$

### 2.3.1 Rappresentazione di uno spazio degli stati discreto attraverso un processo di conteggio

L'idea di base di un processo di conteggio, come il nome forse suggerisce, è di contare il verificarsi di un evento man mano che il tempo procede. Nel nostro caso utilizzeremo un processo di conteggio per contare il verificarsi di tutte le possibili transizioni che possono definirsi all'interno dello spazio degli stati. Utilizzando questo tipo di processi possiamo trasformare uno spazio degli stati astratto in uno completamente equivalente che sia numerico e numerabile.

Definiamo un processo di conteggio univariato  $N(t)$  come un processo in tempo continuo che conta il numero di realizzazioni di un evento nell'intervallo di tempo  $[0, t]$ ; nel nostro caso conterà il numero di transizioni dello stesso tipo.

Formalmente stiamo considerando un processo stocastico a tempo continuo con spazio degli stati nei numeri naturali:

$$\{N(t) \in N : t \in I\}$$

dove  $I$  è un intervallo della retta reale positiva. Il processo di conteggio che si definisce avrà le seguenti proprietà:

i )  $N(t) \in N = \{0, 1, \dots\}$

ii ) con probabilità uno

$$N(t_-) \text{ esiste} \quad e \quad N(t_+) = N(t)$$

i.e. la traiettoria quasi certamente CADLAG

iii )  $N(0) = 0$

iv )  $N(t) - N(t_-) \in \{0, 1\}$

In sostanza abbiamo così definito un processo in tempo continuo e a spazio degli stati discreto con traiettorie non decrescenti e salti unitari negli istanti di tempo nei quali si verifica un evento: si esclude il verificarsi di due eventi dello stesso tipo nello stesso istante.

Per modellizzare la legge di un processo di conteggio possiamo specificare la distribuzione della sequenza degli istanti di salto,  $\tau_k$  e la legge condizionata

$$S_{\tau_k}(t | \tau_1, \dots, \tau_{k-1})$$

oppure la distribuzione della sequenza degli intervalli di tempo che intercorrono tra due salti consecutivi ovvero la distribuzione della durata  $T_k$  con la legge condizionata

$$S_{T_k}(t|T_1, \dots, T_{k-1})$$

Un'altra via naturale, utile per la mia discussione riguardo il ruolo del passato di un processo, è quello di far riferimento alle proprietà dinamiche di un processo e quindi modellizzare la legge distributiva del processo in futuro condizionatamente a quello che è stato il suo sviluppo nel passato, ovvero fino a quel preciso istante.

Definiamo  $\mathcal{F}_N(t)$  la storia del processo  $N(u)$  fino al tempo  $t$  quindi  $\mathcal{F}_N(t)$  è una tribú<sup>9</sup> generata dalla variabile aleatoria  $(N(u) : 0 \leq u \leq t)$ .

L'insieme  $\{\mathcal{F}_N(t) : t \in I\}$  è una famiglia crescente di tribú ovvero una *filtrazione*:

$$\text{se } t' > t \implies \mathcal{F}_N(t') \supset \mathcal{F}_N(t)$$

che rappresenta quindi una accumulazione crescente di informazione. Possiamo considerare  $\mathcal{F}_N(t)$  come la rappresentazione dell'informazione

$$(\tau_1 < \tau_2 < \dots < \tau_p \leq t) \cap \{\tau_{p+1} > t\}^{10}$$

Possiamo ora definire la probabilità istantanea che si verifichi un salto nell'istante di tempo  $t$  condizionatamente al suo passato sviluppo ovvero l'*intensità stocastica* di un processo di conteggio come

$$h_N(t) = \lim_{\epsilon \rightarrow 0} \frac{P[N(t+\epsilon) - N(t) = 1 | \mathcal{F}_N(t_-)]}{\epsilon}$$

dove  $\mathcal{F}_N(t_-)$  è il limite sinistro della filtrazione  $\mathcal{F}_N(t)$  e rappresenta esattamente la storia passata del processo fino al momento esattamente precedente il salto.

Notiamo che l'intensità stocastica è una variabile aleatoria di  $t$  perché dipende da una funzione aleatoria del passato del processo.

Abbiamo introdotto la rappresentazione attraverso i processi puntuali e di conteggio per studiare la transizione di un individuo dallo stato di disoccupato a quello di occupato; resta ora da stabilire il legame tra questi processi e i

<sup>9</sup>Anche  $\sigma$ -algebra  $\mathcal{F}_N(t) = \sigma\{N(u) : 0 \leq u \leq t\}$

<sup>10</sup>Equivalentemente dell'informazione

$$(T_1, T_2, \dots, T_p : \sum_{i=0}^p T_i \leq t) \cap (T_{p+1} > t - \sum_{i=0}^p T_i)$$

modelli di transizione, strumento comunemente usato per descrivere appunto questo tipo cambiamento di stato.

Il modo naturale per la descrizione di un modello di transizione è la determinazione del *tasso di transizione* che rappresenta una probabilità istantanea di accadimento della transizione in un punto  $t$  dato che questa non si è ancora verificata prima di  $t$ ; in termini di un processo di conteggio la stessa informazione è data dalla legge di probabilità condizionata di avere un salto in  $t$  ovvero dall'intensità stocastica.

Se consideriamo un processo semplice i.e. un processo di conteggio con un unico salto è valida la seguente relazione:

$$h_N(t) = h_T(t)\mathbb{I}_{[0,1]}(t) = h_T(t)[1 - N(t_-)]$$

dove l'intensità stocastica del processo,  $h_N(t)$ , viene espressa come funzione del tasso di transizione della durata e il secondo termine a destra dell'uguale indica che il processo è a rischio di salto i.e. la transizione è possibile perché ancora non si è ancora verificata; questo termine rappresenta la storia passata del processo fino a  $t$ . Se consideriamo un processo univariato generale dove è possibile più di una transizione abbiamo

$$h_N(t) = \sum_{k \geq 1} h_{T,k}(t - \tau_k | \tau_1, \tau_2, \dots, \tau_{k-1}) \mathbb{I}_{[\tau_{k-1} < t \leq \tau_k]} = \sum_{k \geq 1} h_{T,k}(T_k | \mathcal{F}_N(t_-))$$

dove l'ultimo termine rappresenta il tasso di transizione per la  $k$ -esima durata condizionato alla storia del processo fino a  $t$ .

Consideriamo ora uno spazio degli stati generale  $E = \{E_0, E_1, E_2, \dots, E_p\}$  ed un insieme,  $\mathcal{C}$ , di tutte le possibili transizioni i.e. tutte le coppie di stati diversi:

$$\mathcal{C} = \{(E_i, E_j) \in E^2 | E_i \neq E_j\}$$

e definiamo per tutte le coppie  $c \in \mathcal{C}$  il processo di conteggio  $N^c(t)$  che conta il numero di transizioni di tipo  $c$  nell'intervallo di tempo  $I = [0, t]$ .

Fissata l'attenzione sulla transizione  $c = (i, j) \in \mathcal{C}$  da  $E_i$  verso  $E_j$  con  $i \neq j$  la sua intensità stocastica sarà

$$h_{ij}(t) = q_{ij}Y_i(t_-)$$

dove  $Y_i(t_-) = \mathbb{I}_{[X(t_-)=E_i]}$ <sup>11</sup> indica se l'individuo è a rischio per la transizione da  $i$  a  $j$  e

$$q_{ij} = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, \mathcal{F}_N(t_-)]}{\epsilon}$$

---

<sup>11</sup>per ogni  $t$  l'informazione data dal processo puntuale  $X(t)$  è equivalente a quella data dal processo di conteggio  $N(t)$  poiché generano lo stesso  $\sigma$ -campo

puó essere vista come una probabilità istantanea per la transizione da  $i$  a  $j$  condizionatamente all'essere nello stato  $E_i$  nell'istante di tempo  $t$  e all'intera storia passata del processo stesso.

Risulta evidente da questa relazione che per modellizzare il processo di transizione tra due stati bisogna modellizzare una probabilità condizionata dipendente dall'intera storia del processo considerato; con l'avanzare del tempo la struttura di questa storia diventa sempre piú complessa e modellizzarla risulta impossibile se non si pongono delle restrizioni al suo effetto sul futuro sviluppo del processo.

### 2.3.2 Caratterizzazione della storia di un processo puntuale

Modellizzare un processo puntuale richiede delle restrizioni sul ruolo della storia passata  $\mathcal{F}_N(t_-)$  sulla legge che governa l'evoluzione nel futuro dello stesso.

Il processo che descrive l'andamento nel passato del processo osservato è caratterizzato da una distribuzione infinito-dimensionale, ma è possibile caratterizzare la sua legge distributiva attraverso delle distribuzioni finito-dimensionali grazie al teorema di Kolmogorov.

Questo teorema mostra che la legge di un processo in tempo continuo è caratterizzato in modo univoco dall'insieme di tutte le distribuzioni marginali multivariate finito-dimensionali del tipo

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \quad \forall t_1 < t_2 < \dots < t_n \quad e \quad n < \infty$$

generate dal processo stesso grazie alla relazione biiettiva tra la legge infinito-dimensionale del processo e questo insieme di distribuzioni finito-dimensionali: chiamiamo questo insieme un sistema proiettivo (projective system). Un projective system è dato da una famiglia finita di distribuzioni definite per ogni  $n < \infty$  che rispettano una condizione di coerenza rispetto alla marginalizzazione; ciò è vero quando dato il sistema finito

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \quad n < \infty$$

e la distribuzione congiunta

$$F_{t_1, t_2, \dots, t_n} = P[X_{t_1} \leq a_1, X_{t_2} \leq a_2, \dots, X_{t_n} \leq a_n]$$

si verifica che

$$\forall n < \infty \quad F_{t_1, t_2, \dots, t_n}(a_1, a_2, \dots, a_{n-1}, \infty) = F_{t_1, t_2, \dots, t_{n-1}}(a_1, a_2, \dots, a_{n-1})$$

In vista di questo teorema possiamo restringere l'influenza della storia del processo caratterizzando il sistema proiettivo del processo in tempo continuo  $X(t)$  in modo da limitare l'influenza di  $\mathcal{F}_N(t_-)$  sullo sviluppo futuro di  $X(t)$ .

Una condizione che restringe fortemente l'influenza del passato è quella di *markovianità* :

il processo puntuale  $X(t) : t \in I$  markoviano se e solo se

$$\forall n \quad \forall t_1 < t_2 < \dots < t_n : X(t_n) \underline{\parallel} (X(t_1), X(t_2) \dots X(t_{n-1})) | X(t_{n-1})$$

ovvero la legge di  $X(t_n)$  condizionatamente a ogni insieme finito delle realizzazioni passate dipende solo da quella piú recente.

In pratica ciò significa che nello studio della transizione è importante solo conoscere lo stato nel quale l'individuo si trovava prima di entrare nel nuovo. Questo tipo di condizione è chiaramente molto restrittiva non tenendo conto, ad esempio, di quanto tempo l'individuo ha trascorso nel vecchio stato: un processo markoviano non ha memoria rispetto alle durate ma solo rispetto agli stati.

L'intesità stocastica della transizione è data in generale da

$$h_{ij}(t) = q_{ij} Y_i(t_-)$$

e sotto l'ipotesi di markovianità si avranno

$$Y_i(t_-) = \mathbb{I}_{[X(t_-)=E_i]}$$

e

$$q_{ij}(t) = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, \mathcal{F}_N(t_-)]}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i]}{\epsilon}$$

Una classe di processi piú generali si definisce attraverso la propriet di *semi-markovianità* :

il processo puntuale  $X(t) : t \in I$  è semi-markoviano se e solo se

$$\forall n \quad \forall t_1 < t_2 < \dots < t_n : (T_n, X(t_n)) \underline{\parallel} (T_1^{n-1}, X_1^{n-1}) | X(t_{n-1})$$

dove  $X_1^{n-1}$  la sequenza degli stati e  $T_1^{n-1}$  la sequenza delle durate ad essi associate. in un processo semi-markoviano non si considera solo il vecchio stato ma anche quanto tempo l'individuo lo ha occupato. <sup>12</sup>.

<sup>12</sup>Notiamo che in un processo semi-markoviano il sotto-processo generato dalla sequenza degli stati visitati forma un processo markoviano . La differenza riguarda il ruolo del processo di durata che è distribuito esponenzialmente in un processo markoviano (assenza di memoria) ed ha una distribuzione arbitraria  $F$  in un processo semi-markoviano.

L'intensità stocastica di una transizione per un processo semi-markoviano sarà caratterizzata da:

$$Y_i(t_-) = \mathbb{I}_{[X(t_-)=E_i]}$$

e

$$q_{ij}(t) = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, \mathcal{F}_N(t_-)]}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, T_i]}{\epsilon}$$

Abbiamo visto che la storia di un processo puntuale  $X(t) : t \in I$  è data dalla filtrazione

$$\begin{aligned} \mathcal{F}_X(t_-) &= \sigma\{(\tau_1, X(\tau_1), \dots, (\tau_k, X(\tau_k)), \mathbb{I}_{[\tau_{k+1} > t]}\} \\ &= \sigma\{T_1, X(T_1), \dots, T_k, X\left(\sum_{0 \leq i \leq k} T_i\right), \mathbb{I}_{[\sum_{i=0}^{k+1} T_i > t]}\} \end{aligned}$$

dove  $k$  è una funzione del tempo :  $k(t) = \max\{j | j \leq t\}$  quindi senza nessuna restrizione l'intensità stocastica del processo ,  $h_X(t)$  dipenderá da  $2k(t) + 1$  variabili aleatorie.

Il ruolo della proprietà markoviana è quello di ridurre questo numero a 1 (solo  $X(\tau_{k-1})$ ) e quella semi-markoviana a 2 (  $X(\tau_{k-1})$  e  $(t - \tau_{k-1})$  ).

### 2.3.3 Alcune considerazioni riguardo processi piú generali

In alcuni casi le condizioni di markovianità o semi-markovianità possono risultare troppo restrittive perché si vuole prendere in considerazione una parte piú lunga della storia del processo. Potrebbe essere piú soddisfacente richiedere all'intensità stocastica del processo di dipendere da un numero finito  $p > 2$  di variabili, con  $p$  indipendente da  $t$ ; questa indipendenza permette di utilizzare per ogni istante di tempo  $t$  lo stesso numero di variabili come una sorta di sunto dell'intera storia fino all'istante esattamente precedente  $t$ . Quello che immaginiamo è l'esistenza di un vettore  $W(t) = (W_1(t), W_2(t), \dots, W_p(t))$  dove ogni componente,  $W_j(t)$ , dipende dalla storia passata ,  $\mathcal{F}_X(t)$ , e tale che l'intensità stocastica del processo dipenda solo da tale vettore :

$$h_X(t) = f(\tau_1, X(\tau_1), \dots, \tau_k, X(\tau_k), \mathbb{I}_{[\tau_{k+1} > t]})$$

diventa

$$h_X(t) = f(W(t)) = \{(W_1(t), W_2(t), \dots, W_p(t))\}$$

quello che si cerca è una famiglia di  $\sigma$ -campi  $\mathcal{W}(t)$  :

$$\mathcal{W}(t) = \sigma\{E[f(X(t))|\mathcal{F}_X(t-)] : f \in L_1(\mathcal{F}_X(t-))\}$$

che renda possibile la scrittura dell'intensità stocastica come

$$h_X(t) = f(\mathcal{F}_X(t-)) = f(\mathcal{W}(t))$$

Il punto importante è rappresentato dal fatto che  $\mathcal{W}(t)$  non è una filtrazione e quindi non è una accumulazione crescente di informazioni ma una sorta di riassunto, dipendente da un numero fisso di variabili, di tale filtrazione .

$\mathcal{W}(t)$  una famiglia di  $\sigma$ -campi tale che:

- $\mathcal{W}(t)$  adattata a  $\mathcal{F}_X(t-)$ , *i.e.*  $\mathcal{W}(t) \subset \mathcal{F}_X(t-)$
- $\mathcal{W}(t)$  un predittore sufficiente per la legge del processo  $X(t)$   
*i.e.*  $X(t) \parallel \mathcal{F}_X(t-)|X(t_0), \mathcal{W}(t)$

Con queste due condizioni si ha una funzione dell'intera storia del processo che dá *solo* le informazioni sufficienti alla predizione e conoscenza della legge di  $X(t)$  e non tutte quelle realmente costituenti questo passato.

Questo ci permette di stimare un tasso di transizione, funzione dell'intensità stocastica, senza considerare l'intera storia dell'individuo ma nemmeno una parte molto ristretta di questa, bensí una funzione dipendente da un numero finito di parametri completamente determinata dall'intero passato fino all'istante della transizione.

Anche questa funzione dipenderá dal tempo ma in un senso diverso dalla filtrazione.

In una filtrazione il tempo produce variabili aleatorie che si vanno ad aggiungere alla sequenza di quelle già esistenti fino  $t-$  : una filtrazione è una sequenza crescente di variabili aleatorie.

Nel caso della funzione  $\mathcal{W}(t)$  il tempo ha la funzione di far cambiare il valore delle componenti, o di alcune di loro, del vettore  $W(t)$  a causa del cambiamento nelle caratteristiche del processo che stiamo utilizzando per riassumere il passato.

Queste considerazioni ci portano ad un'altra proprietà che richiediamo per la funzione  $W(t)$  che è quella della sua *adattamento* con il tempo: ogni volta che una nuova variabile entra nel processo la storia si modificherá implicando anche una modificazione dei valori di questo vettore.

In altre parole richiediamo che:

$$\mathcal{W}(t+1) = g[\mathcal{W}(t), I_{t+1}]$$



dove  $I_{t+1}$  rappresenta l'informazione disponibile al tempo  $t$ .  
 In questo contesto l'intensit  stocastica del processo sar 

$$h_{ij}(t) = q_{ij}(t)Y_i(t_-)$$

con  $Y_i(t_-) = \mathbb{I}_{[X(t_-)=E_i]}$  e

$$q_{ij}(t) = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, \mathcal{F}_N(t_-)]}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{P[X_{t+\epsilon} = E_j | X_t = E_i, \mathcal{W}(t)]}{\epsilon}$$

In questo lavoro verr  stimato un modello sotto l'ipotesi che per avere una buona conoscenza del processo in analisi, ovvero la durata della disoccupazione, sia sufficiente ricordare solo alcune caratteristiche del passato dell'individuo rispetto alla disoccupazione e non l'intera storia e quindi che i due modelli condizionali

$$X_t | \mathcal{F}_X(t_-) \quad e \quad X_t | \mathcal{W}(t)$$

sono equivalenti in termini di informazioni sulla legge di  $X_t$ .

Un'altra questione importante   quella di confrontare questo modello con uno stimato sotto l'ipotesi di semi-markovianit  e verificare se effettivamente c'  un incremento nelle capacit  esplicative nei confronti del processo; potrebbe infatti verificarsi che per un processo sia inutile appesantire il modello esplicativo con una funzione del passato poich  questo non ha nessuna influenza e quindi il processo stesso sottost  ad ipotesi semi-markoviane o markoviane.

## 2.4 Definizione e scelta di un modello parametrico

Quando si lavora con i modelli di transizione   molto importante la definizione dello spazio degli stati in quanto uno spazio degli stati non corretto ovvero non esaustivo pu  portare a forti distorsioni.

Heckman e Singer [Heckman,Singer,1984b] (cap.6) definiscono il *basic unemployment model* attraverso 3 stati: occupato, disoccupato, out-of-labor-force. Viene ampiamente discussa l'importanza della distinzione tra chi   all'interno del mercato del lavoro e disoccupato e chi invece pu  essere considerato tale ma in realt  non lo   in quanto fuori dal mercato del lavoro come ad esempio le casalinghe [Flinn,Heckmann,1983]. Essendo individui comportamentalmente molto differenti considerarli assieme porterebbe ad una forte distorsione nell'analisi della condizione di disoccupato. Questi tre stati sono stati distinti anche per strutturare il modello utilizzato in questo lavoro, l'attenzione   stata poi rivolta alla sola transizione disoccupazione-occupazione

ma il terzo stato é servito a distinguere quegli episodi di inattività che non possono però considerarsi di disoccupazione. Gli stati verranno definiti esattamente nel cap.3 dedicato alla struttura e al modo in cui sono stati elaborati i dati del Panel Europeo.

I dati a disposizione coprono un periodo di 5 anni e ciò fa sì che ogni individuo, potenzialmente, vive più episodi di disoccupazione; volendo studiare la transizione all'occupazione si pone il problema di come modellizzare una struttura a episodi multipli.

Quando si considera un insieme di episodi singoli si considerano tali episodi statisticamente indipendenti e provenienti da una popolazione omogenea; in un set di episodi multipli, gli individui con più episodi sono rappresentati maggiormente e gli episodi dello stesso individuo non sono indipendenti. In quest'ultimo caso, in fase di modellizzazione, si possono effettuare almeno tre scelte [Blossfeld, Rohwer, 1995] p.79:

- 1 ) Considerare solo gli episodi della stessa cardinalità ovvero fare l'analisi per sotto-gruppi.
- 2 ) Considerare gli episodi come indipendenti condizionatamente ad un set di covariate che contenga informazioni rilevanti della loro storia e considerare esplicitamente l'eterogeneità della popolazione.
- 3 ) Esplicitare una struttura multi-episodio e multi-transizione e lavorare su specifiche traiettorie [Flinn, Heckmann, 1982].

In questo studio verrà seguita la seconda ipotesi essenzialmente per due motivi: innanzitutto, la selezione degli episodi in base alla cardinalità non porterebbe, a mio avviso, a gruppi di significato. I primi episodi andrebbero a definire il gruppo dei primi episodi osservati nel quale comunque confluiscono episodi di cardinalità diversa in quanto gli individui non sono osservati, ad esempio, dal loro primo ingresso nel mercato del lavoro ma in un momento qualsiasi della loro storia lavorativa. Non potendo identificare temporalmente e relativamente ad un punto ben preciso della storia di vita di ogni individuo la cardinalità vera dell'episodio osservato la definizione dei gruppi attraverso essa avrebbe reso poco interpretabili i risultati. Inoltre la numerosità dei sotto-gruppi che si possono definire potrebbe essere molto diversa e ancor più diverso il numero delle transizioni per ogni sotto-gruppo: più é piccolo il numero delle transizioni per una determinata cardinalità più facilmente si possono ottenere stime statisticamente non significative. Il confronto stesso dei coefficienti significativi sulla stessa variabile ma per cardinalità diversa potrebbe, per lo stesso motivo, portare a conclusioni distorte. In secondo luogo, la variabile relativa alla cardinalità dell'episodio rappresenta il primo

passo per la costruzione della storia del processo che si vuole analizzare. La variabile rappresentante la cardinalità dell'episodio permette da un lato di controllare la dipendenza tra gli episodi dello stesso individuo e dall'altro di esplorare l'effetto sull'episodio  $n$  degli  $n - 1$  episodi precedenti. Ovviamente la cardinalità dell'episodio è solo il primo passo per questa esplorazione e per la costruzione della funzione della storia passata del processo  $W(t)$ .

Una volta definiti gli episodi si passa alla scelta del modello.

Desiderando utilizzare un modello parametrico si procederà con il testare quanto funzioni di rischio derivate da forme parametriche diverse si adattino ai dati. Le funzioni di rischio parametriche, con parametri stimati a partire dai dati, vengono confrontate con la funzione di rischio stimata non parametricamente sullo stesso campione attraverso test di adattamento. I test che si possono utilizzare si dividono essenzialmente in due categorie: test grafici e test che si basano sugli pseudoresidui.

### 2.4.1 Test grafici

Questo tipo di test si basa su trasformazioni della funzione di sopravvivenza stimata sotto l'ipotesi che la funzione parametrica che si sta utilizzando sia quella adatta ai dati; una volta scelta la distribuzione da dare alla variabile di durata la funzione di sopravvivenza resta univocamente determinata a livello analitico e sarà funzione del tempo trascorso nello stato del quale si sta analizzando la transizione. Una volta determinata la forma analitica si effettuano trasformazioni che permettono di rileggerla come funzione lineare del tempo o di una sua trasformata. Il plot della funzione di sopravvivenza trasformata e della trasformazione del tempo risulta perfettamente lineare se la forma parametrica scelta si adatta perfettamente ai dati: minore è l'adattamento e più la curva risultante si allontanerà da una retta. Questo permette anche di stimare un modello di regressione lineare tra le due trasformazioni e vedere quanto la curva rappresentata si allontana da essa nonché di ottenere l'indice di determinazione lineare da usare come misura della dipendenza lineare della trasformata della funzione di sopravvivenza dalla trasformata del tempo. Per esemplificare meglio questo procedimento vediamo come si applica su alcune funzioni tra le più utilizzate per la costruzione di modelli di durata parametrici.

#### Distribuzione log-logistica

La funzione di densità di una variabile aleatoria  $T$  con distribuzione log-logistica è data da:

$$f_T(t) = \frac{\beta \alpha t^{\beta-1}}{(\alpha t^\beta + 1)^2}$$

da cui risulta che la funzione di sopravvivenza é data da:

$$S_T(t) = \frac{1}{\alpha t^\beta + 1}$$

Notiamo che  $1 - S_T(t) = \frac{\alpha t^\beta}{1 + \alpha t^\beta} \Rightarrow \frac{1 - S_T(t)}{S_T(t)} = \alpha t^\beta \Rightarrow \log\left(\frac{1 - S_T(t)}{S_T(t)}\right) = \log(\alpha) + \beta \log(t)$ .

L'ultima equazione mostra l'esistenza di una relazione lineare tra  $\log\left(\frac{1 - S_T(t)}{S_T(t)}\right)$  e  $\log(t)$  ovvero che un plot dovrebbe dare approssimativamente una retta. La stima di un modello di regressione in cui la trasformata della funzione di sopravvivenza viene fatta dipendere da  $\log(t)$  ci permette di ottenere delle stime per i parametri  $\log(\alpha)$  e  $\beta$  e quindi di sovrapporre la retta cosí ottenuta al plot precedente per valutarne la *somiglianza* graficamente.

### Distribuzione log-normale

La funzione di densitá di una variabile aleatoria  $T$  distribuita come una log-normale di parametri  $\mu$  e  $\sigma$  é data da:

$$f_T(t) = \frac{1}{\sigma t \sqrt{2\pi}} \exp^{-\frac{1}{2\sigma^2}(\ln t - \mu)^2} = \frac{1}{\sigma t} \varphi\left(\frac{\ln t - \mu}{\sigma}\right)$$

dove  $\varphi\left(\frac{\ln t - \mu}{\sigma}\right)$  rappresenta la densitá di una variabile aleatoria con distribuzione normale standardizzata. La variabile in questo caso é  $\ln t$  infatti

$$T \sim LN(\mu, \sigma) \iff \ln T \sim N(\mu, \sigma)$$

Dalla densitá cosí definita risulta una funzione di sopravvivenza data da:

$$S_T(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

con  $\Phi$  funzione di distribuzione di una variabile normale standardizzata. Identificando con  $\Phi^{-1}$  e l'inversa di questa distribuzione si ha che

$$\Phi^{-1}(1 - S_T(t)) = -\frac{\mu}{\sigma} + \frac{1}{\sigma} \ln t$$

se il modello log-normale é quello appropriato un plot di  $\Phi^{-1}(1 - S_T(t))$  versus  $\ln t$  dovrebbe essere approssimativamente lineare. Come visto per la distribuzione log-logistica, anche in questo caso la trasformata della funzione di sopravvivenza e della variabile temporale possono essere messe in relazione con un modello di regressione semplice in grado di fornire una stima per i parametri della relazione lineare introdotta sopra.

Per le trasformazioni da utilizzare per verificare l'adattamento grafico di altre distribuzioni di veda ad esempio [Lawless,1982] , [Mouchart,2003].

## 2.4.2 L'analisi degli pseudoresidui

Informazioni aggiuntive a quelle fornite dai test grafici sulla scelta di un modello parametrico si possono ottenere attraverso l'analisi degli *pseudoresidui* o *residui generalizzati*. Nelle analisi con modelli di regressione stimati attraverso i minimi quadrati ordinari (OLS) il metodo tradizionale e forse migliore di valutare le possibili violazioni delle ipotesi sottostanti il modello é attraverso una stima diretta dei residui. Questi rappresentano quindi deviazioni tra i valori osservati della variabile dipendente e quelli stimati sotto le ipotesi di un modello specifico. Nei modelli di durata la variabile dipendente é il tasso di transizione che é però non osservabile. Ciò rende impossibile la stima diretta dei residui come risultato della differenza tra valori osservati e stimati per ogni unità.

Per i modelli di transizione esiste in ogni caso un approccio simile basato sulla definizione degli pseudoresidui introdotta da Cox e Snell nel 1968 [Cox,Snell,1968].

Definiamo  $\hat{h}_T(t; \mathbf{x})$  lo stimatore della funzione di rischio basata su un campione di individui  $i = 1, \dots, N$  con durate osservate  $t_i$  e vettore di covariate  $x_i$ . Gli pseudoresidui sono definiti come tasso di transizione cumulato, valutato per quel dato campione, ovvero:

$$\hat{e}_i = \int_0^{t_i} \hat{h}_T(\tau_i; x_i) d\tau \quad i = 1, \dots, N$$

La motivazione sottostante la definizione é che se il modello é appropriato e non ci sono durate censurate i residui così definiti risultano non informativi ovvero non danno un contributo significativo alla definizione della forma della funzione di rischio. I residui sono non informativi quando la loro distribuzione definisce la costanza del rischio ovvero quando la variabile di durata, che in questo caso sarà rappresentata dai residui, segue una distribuzione esponenziale. Se non ci sono episodi censurati e il modello é appropriato gli pseudoresidui risulteranno essere un campione proveniente da una distribuzione esponenziale standard. Se ci sono dati censurati allora anche i corrispondenti residui saranno guardati come censurati. In entrambi i casi é possibile ottenere una stima, ad esempio con il metodo di Kaplan-Meyer, della funzione di sopravvivenza dei residui  $S_{\hat{e}}(\hat{e})$ . Se i residui seguono una distribuzione esponenziale allora  $S_{\hat{e}}(\hat{e}_i) = \exp(-\alpha \hat{e}_i)$  con  $h_{\hat{e}}(\hat{e}_i) = \alpha$ .

La trasformata logaritmica della funzione di sopravvivenza evidenzia una relazione lineare e quindi si può utilizzare un test grafico sul plot di  $-\ln(S_{\hat{e}}(\hat{e}_i))$  versus  $\hat{e}_i$  per vedere se i residui seguono una distribuzione esponenziale. L'idea é quindi quella di trattare gli pseudoresidui come durate spurie che possono accettarsi in quanto non influenzano la relazione temporale osservata. Con questa procedura si torna in ogni caso ad una valutazione di carattere

grafico e quindi questo utilizzo degli pseudoresidui non può essere considerato a tutti gli effetti un test di valutazione della bontà di adattamento di un modello ad un insieme di dati quanto una buona indicazione rispetto alla direzione da prendere: ad esempio permette di discriminare bene tra l'adattamento di un rischio monotono piuttosto che campanulare.

## 2.5 L'eterogeneità non osservata

Utilizzando i modelli di transizione si cerca di trovare evidenza su come il tasso di transizione, che descrive il movimento degli individui in un determinato spazio degli stati, dipende da un insieme di covariate. Sfortunatamente, non sempre si è in grado di includere tutti i fattori importanti, perché non si hanno i dati a disposizione o semplicemente perché non si possono conoscere a priori. Affrontando questa situazione bisogna tener presenti almeno due aspetti importanti.

La prima considerazione è ben nota quando si parla di modelli di regressione tradizionali. Le covariate incluse in un modello sono normalmente correlate e la stima dei parametri dipende dall'insieme specifico di variabili utilizzato: ogni cambiamento in questo insieme è capace di modificare le stime dei parametri di variabili incluse precedentemente.

In secondo luogo, quando si tratta di modelli di transizione, bisogna ricordare che cambiare l'insieme delle covariate può fortemente modificare la dipendenza temporale del tasso di transizione. Un effetto simile si osserva anche nei modelli di regressione tradizionali in quanto a seconda dell'insieme di covariate utilizzato, la distribuzione empirica dei residui si modifica. Ma, contrariamente ai modelli di regressione tradizionali dove i residui sono utilizzati normalmente solo per testare le ipotesi del modello, nei modelli di transizione i residui diventano l'oggetto del modello.

Se i modelli di transizione vengono riformulati come modelli di regressione, il tasso di transizione diventa una descrizione dei residui e ogni cambiamento nella distribuzione dei residui determina un cambiamento nella forma della tempo dipendenza del tasso di rischio [Blossfeld,Rohwer,1995], [Heckman,Singer,1984a], [Lancaster,1990], [Lawless,1982].

La questione diventa quindi se e in che senso un modello di transizione può fornire almeno qualche verosimile conoscenza sulla tempo dipendenza del rischio dovendo considerare l'esistenza di una eterogeneità non osservabile dovuta alla diversità di essere e di esperienze delle unità del campione.

Come trattare l'eterogeneità non osservata in un modello di transizione? Come per i modelli di regressione classici è possibile incorporare un termine di errore nel modello di transizione anche se le conseguenze e considerazioni

a riguardo sono differenti.

Consideriamo una formulazione generale dei modelli di transizione come

$$h_T(t) = f(x, \beta, \epsilon)$$

dove il tasso di transizione dipende da un vettore di covariate e dai loro parametri nonché da una variabile aleatoria che rappresenta un termine di errore. Per poter stimare il modello è necessario formulare ipotesi sulla relazione funzionale  $f(\cdot)$  e sulla forma distributiva di  $\epsilon$ .

Scriviamo il modello di transizione sotto forma di modello di regressione [Lawless,1982] come

$$\ln T = x\beta + \epsilon$$

A seconda della distribuzione che si utilizzerà per  $\epsilon$  si otterrà un modello di transizione. Ad esempio, assumendo che  $\epsilon$  sia distribuito normalmente,  $\epsilon \sim N(0, \sigma^2)$ , la variabile di durata  $T$ , condizionatamente alle covariate utilizzate, seguirà una distribuzione log-normale con densità data da

$$\frac{1}{\sigma t} \Phi\left(\frac{\ln T - \beta}{\sigma}\right)$$

ovvero si otterrà un modello di transizione con specificazione log-normale del rischio di base. L'assunzione sulla distribuzione del termine di errore si risolve quindi in un'assunzione sulla distribuzione della variabile di durata  $T$ . È come se il modello contenesse già un termine di errore: la distribuzione di  $T$  condizionata alle covariate.

Questa formulazione non può considerarsi del tutto corretta in quanto nei modelli di regressione si modella il valore atteso della variabile dipendente mentre nei modelli di transizione il focus è la distribuzione di  $T$  e non il suo valore atteso. Anche se da un punto di vista statistico non ci sono grosse differenze tra i due approcci alla modellizzazione ve ne sono in termini teorici in quanto il tasso di transizione ha un significato sostanziale insito nel suo rappresentare la *propensione* degli individui a cambiare stato.

L'approccio utilizzato con i modelli di transizione è quello di separare il termine di errore  $\epsilon$  in due componenti; il modello sotto forma regressiva diventerà

$$\ln T = x\beta + \epsilon_1 + \epsilon_2$$

Possiamo ora specificare una forma distributiva per  $\epsilon_1$  e quindi per la durata  $T$  per definire così il modello di transizione che conterrà il termine di errore  $\epsilon_2$ . Per rendere il modello stimabile ci sarà bisogno di ipotesi anche sulla distribuzione di quest'ultima componente. Facendo delle ipotesi su  $\epsilon_1$  e  $\epsilon_2$  abbiamo in sostanza fatto ipotesi su  $\epsilon = \epsilon_1 + \epsilon_2$ ; si è cioè specificata una

distribuzione mista per  $\epsilon$  e di conseguenza per  $T$  e ancora una volta si ha un modello di transizione ma senza termine di errore.

Il fatto essenziale è che separare il termine di errore in due parti è una procedura assolutamente arbitraria; da un punto di vista empirico abbiamo un solo insieme di residui dovuti alla specificazione del modello e sarebbe completamente arbitrario separarli in due componenti, una da interpretarsi come descrizione del tasso di transizione e l'altra come termine di errore. Anche se questi modelli non risultano utili per distinguere la vera forma del rischio dal termine di errore arricchiscono notevolmente lo spettro di modelli che si possono utilizzare per arrivare a dei risultati robusti.

Per stimare modelli di transizione che si basano su una mistura di distribuzioni si possono seguire essenzialmente due approcci: utilizzare una specificazione completamente parametrica della distribuzione del termine di errore oppure basarsi su una distribuzione di mistura discreta che può ritenersi un approccio semi-parametrico in quanto richiede ipotesi molto deboli. Di seguito verrà esplicitato il primo approccio in quanto utilizzato poi in fase empirica.

### 2.5.1 Modelli con mistura di tipo Gamma

L'assunzione sottostante l'utilizzo di un termine di eterogeneità nei modelli di transizione è che vi sia una costante indipendente dal tempo e specifica per ogni unità d'analisi; si assume che tale costante non osservata sia la realizzazione di una variabile casuale identicamente distribuita per ogni unità e indipendente dalle covariate osservate. Tale termine è legato moltiplicativamente al tasso di transizione con l'idea di renderlo capace di cogliere la presenza di eterogeneità non osservata ma anche di non rispetto della forma parametrica utilizzata nella specificazione del rischio di base.

In sostanza si assume che il tasso di transizione,  $h_T(t|x, \nu)$ , dipenda da un insieme di covariate,  $x$ , e da un termine stocastico  $\nu$  non osservato <sup>13</sup>.

L'ipotesi di moltiplicatività dell'effetto di  $\nu$  porta a scrivere la funzione di rischio come

$$h_T(t|x, \nu) = h_T^b(t|x)\nu \quad \nu \geq 0$$

con

$$H_T(t|x, \nu) = \nu H_T^b(t|x) = \nu \int_0^t h_T^b(\tau|x) d\tau$$

Con questa espressione si assume che la funzione di rischio sia determinata dalla funzione di rischio di base moltiplicata per uno scalare  $\nu$ .

La seconda ipotesi che facciamo è che  $\nu$  segua una distribuzione di tipo

---

<sup>13</sup>Questa variabile aleatoria corrisponde al termine di errore  $\epsilon_2$



Gamma con valore atteso unitario,  $E(\nu) = 1$ ; da ciò deriva che la funzione di densità di tale variabile è data da

$$f_\nu(\nu) = \frac{k^k \nu^{k-1}}{\Gamma(k)} \exp(-k\nu) \quad k \geq 0$$

La varianza di tale distribuzione è data  $Var(\nu) = 1/k$ .

Il passo successivo è calcolare la densità della mistura risultante e derivare la funzione di rischio da stimare.

Per le funzioni di densità e sopravvivenza abbiamo

$$f_T(t|x, \nu) = \int_0^\infty f_T(t|x, \nu) f_\nu(\nu) d\nu$$

$$S_T(t|x, \nu) = \int_0^\infty S_T(t|x, \nu) f_\nu(\nu) d\nu$$

che rappresentano valori attesi rispetto alla distribuzione di  $\nu$  facili da calcolare poiché per la funzione Gamma vale la seguente relazione

$$\int_0^\infty \exp(-\nu g(t)) f_\nu(\nu) d\nu = \left[1 + \frac{1}{k} g(t)\right]^{-k}$$

Sulla base di questa relazione si ottengono densità e sopravvivenza non condizionati per l'insieme di episodi ovvero

$$S_T(t|x) = \int_0^\infty \exp(-\nu H_T^b(t|x)) f_\nu(\nu) d\nu = \left[1 + \frac{1}{k} H_T^b(t|x)\right]^{-k}$$

$$f_T(t|x) = h_T^b(t|x) \left[1 + \frac{1}{k} H_T^b(t|x)\right]^{-k-1}$$

da cui si ricava il tasso di rischio non condizionato come

$$h_T(t|x) = \frac{f_T(t|x)}{S_T(t|x)} = h_T^b(t|x) \left[1 + \frac{1}{k} H_T^b(t|x)\right]^{-1}$$

i parametri di questa funzione verranno stimati con il metodo della massima verosimiglianza utilizzando le funzioni di densità e sopravvivenza non condizionate risultanti dal modello di mistura.



## Capitolo 3

# Dati utilizzati e loro organizzazione

### 3.1 L'European Community Household Panel (ECHP)

L'ECHP è un data base longitudinale risultato di un progetto lanciato nel 1992 dall'Eurostat in risposta alla crescente domanda di informazioni comparabili tra gli stati membri dell'Unione Europea riguardanti reddito, lavoro e occupazione, povertà ed esclusione sociale, condizioni abitative, salute e tanti altri indicatori sociali riguardanti le condizioni di vita di famiglie ed individui.

Le indagini hanno cadenza annuale e sono svolte dalle National Data Collection Units (NDU) rappresentate dagli istituti nazionali di statistica, come in Italia dove è l'ISTAT ad occuparsi della rilevazione, o da altre organizzazioni, comunque coordinate e supportate dall'Eurostat.

La popolazione di riferimento consiste di tutte le famiglie residenti nel territorio nazionale; l'estrazione del campione è affidata alle NDU e possono essere differenti pur rispettando le linee guida generali fornite da Eurostat per assicurare la loro comparabilità. La procedura campionaria più utilizzata prevede un campionamento a due stadi con il primo definito attraverso le aree geografiche e il secondo attraverso la lista delle famiglie (o degli indirizzi) residenti nelle sotto aree definite. All'interno di ogni famiglia verranno intervistati tutti componenti adulti ovvero che hanno più di 16 anni al 31 dicembre dell'anno precedente quello dell'indagine. L'Eurostat ha messo a punto una procedura di calcolo di pesi a livello individuale e a livello familiare che rendono i dati ECHP rappresentativi della popolazione di riferimento per ogni anno: i dati annuali possono essere utilizzati come un campione cross-

section.

Le famiglie che vengono seguite sono quelle intervistate all'ondata  $t - 1$  e quelle non intervistate per mancanza di contatto, indisposizione, problemi fisici o psichici nonché ogni nuova famiglia contenente almeno un individuo campionato. Sono escluse dal follow-up invece le famiglie che si trasferiscono in uno stato esterno all'Unione Europea, che avevano dato rifiuto finale all'intervista precedente o che sono state completamente istituzionalizzate. Dalla terza ondata in poi vengono comunque escluse quelle che non sono state intervistate per due ondate consecutive. Per la maggior parte degli stati partecipanti all'ECHP il declino dell'ampiezza del campione delle famiglie è moderato e per alcuni, tra cui l'Italia e il Portogallo, si osserva un leggero incremento [Peracchi, Nicoletti, 2002].

I dati vengono raccolti attraverso un questionario standard elaborato dall'Eurostat consistente in una parte da somministrare ad una persona di riferimento atta a raccogliere le informazioni a livello familiare e un'altra parte da somministrare a tutti gli adulti della famiglia per raccogliere le informazioni a livello personale. A grandi linee le aree indagate sono:

- Migrazioni e abitazione; tempo trascorso nelle diverse abitazioni, ragioni per cambiare abitazione, tipo e caratteristiche dell'abitazione attuale etc...
- Situazione economica; se e come riesce a procurarsi determinati beni, fare fronte a diverse necessità etc...
- Reddito nell'anno di calendario precedente l'intervista; ammontare, fonte etc..
- Condizioni di salute e rapporti sociali.

Il questionario personale risulta ovviamente più corposo e ricco di particolari riguardo all'attività lavorativa e alle strategie utilizzate per inserirsi nel mercato del lavoro.

Il periodo di riferimento varia considerevolmente a seconda dell'oggetto sul quale si sta indagando. Per le caratteristiche personali è il giorno dell'intervista, per le attività economiche (ad esempio lo stato rispetto all'occupazione) sono gli ultimi sette giorni precedenti l'intervista: per chi non lavora il periodo di riferimento è dato dalle ultime quattro settimane lavorative prima dell'intervista. Per i redditi ci si riferisce all'ultimo anno di calendario mentre

per la situazione finanziaria e dell'attività in genere ad ogni mese dell'anno precedente l'intervista.

Il metodo di intervista consigliato da Eurostat è l'intervista faccia-a-faccia anche se in alcuni stati è stata utilizzata ad esempio la tecnica CAPI (Computer Assisted Personal Interviewing); la compilazione dei questionari e l'inserimento dei dati in forma elettronica porta alla creazione del Production Data Base (PDB). Siccome il PDB contiene informazioni ritenute confidenziali sulla base delle leggi statistiche dell'Unione Europea ed ha una struttura molto complessa il suo accesso è riservato alle NDU e ad Eurostat; per gli utilizzatori esterni viene fornito lo User's Data Base (UDB) che è una versione anonimizzata e user-friendly del PDB.

L'UDB consiste di quattro file per ogni ondata: un *Register file*, un *Relationship file*, un *Household file* e un *Personal file*. Il *Register file* rappresenta l'elenco di tutte le persone che vivono in una famiglia per la quale si è avuta un'intervista completa. Il *Relationship file* specifica le relazioni tra i membri della famiglia e contiene un record per ogni coppia di persone nella famiglia. L'*Household file* contiene i dati delle interviste familiari e quindi un record per ogni famiglia mentre il *Personal file* quelli provenienti dalle interviste personali e quindi un record per ogni individuo. Le famiglie e i loro componenti sono individuate attraverso un numero di identificazione che è fisso e unico all'interno di ogni stato partecipante. L'UDB contiene anche un *link file* che permette di ricostruire la traiettoria degli individui tra le diverse ondate e un *country file* con indicatori a livello nazionale come l'ammontare di popolazione o il tasso di cambio.

L'EHCPC risente di tutti i problemi tipici delle indagini campionarie nonché panel come le mancate risposte parziali o totali e l'attrito. Per una valutazione dei loro effetti sui dati italiani e un confronto con la situazione degli altri stati partecipanti all'indagine si veda [Peracchi,2000] pag.73 e [Peracchi,Nicoletti,2002].

## 3.2 Organizzazione dei dati

La popolazione di riferimento per questo lavoro è data dagli individui con età compresa tra 16 e 35 anni. Questi sono stati selezionati dal *Personal file* della prima ondata e seguiti successivamente sulle altre in modo da poter ricostruire il loro percorso attraverso i cinque anni osservati. Persone che entrano nel panel in periodi successivi e della stessa fascia di età entrano comunque a far parte della popolazione in studio.

La sezione che ha costituito la base di partenza per la costruzione della matrice sulla quale stimare il modello è data dalla sezione denominata **Calen-**

**dario delle attività** , vedere l'appendice A per la struttura e le informazioni contenute nei file individuali.

Questa sezione riporta l'attività principale dell'individuo per ogni mese dell'anno precedente quello di indagine; è stato quindi possibile ricostruire l'attività principale per ogni mese dal gennaio 1993 al dicembre 1997. Le altre informazioni contenute nel file si riferiscono però all'anno in corso e quindi l'attività principale è stata ricostruita solo per gli anni dal 1994 al 1997 partendo dai file dal 1995 al 1998 in modo da poter riferire gli eventi vissuti alle caratteristiche possedute in quell'anno e non a quelle possedute nell'anno successivo all'evento. Ciò porta alla definizione di un intervallo temporale  $I = [0, 47]$  con 0=gennaio 1995 e 47=dicembre 1997.

L'attività principale viene codificata attraverso 10 categorie che sono state riaggregate per arrivare alla definizione di tre stati: occupato , disoccupato e fuori dal mercato del lavoro. Per i significati delle 10 categorie iniziali e le corrispondenze con i tre stati finali si veda Appendice B.

Il primo stato comprende quegli individui che dichiarano di lavorare a tempo pieno o parziale presso una ditta ovvero di lavorare all'interno di qualche progetto di formazione, di essere lavoratore autonomo o di lavorare presso l'impresa familiare (anche se senza stipendio). Lo stato di disoccupato contiene solo quegli individui che si identificano come tali mentre lo stato che identifica quelli al di fuori del mercato del lavoro comprende studenti, casalinghe, soldati e qualsiasi altra forma di inattività non accompagnata dalla ricerca di lavoro.

Conoscendo lo stato occupato dagli individui per ogni mese è possibile costruire i dati che rappresentano il processo di durata ovvero la sequenza degli stati attraversati accompagnata dalle rispettive durate. Il risultato sarà una matrice dove ogni riga rappresenta un episodio e registra i momenti di inizio e fine, quindi la sua lunghezza, nonché il vecchio e il nuovo stato che permettono l'identificazione del tipo di transizione per cui viene chiamata **matrice di transizione** .

Avendo definito tre stati sono infatti possibili sei tipi di transizioni: non è possibile la transizione diretta da uno stato verso se stesso. Gli episodi generati dallo stesso individuo vengono identificati attraverso il numero di identificazione personale che ci permette di creare una struttura multi-episodio con il giusto riferimento sia all'individuo che temporale. Questa operazione costituisce il passaggio da un *individual-oriented data set* ad un *episode-oriented data set* in cui l'unità di analisi diventa l'episodio osservato a cui vengono associate le caratteristiche dell'individuo che l'ha prodotto. Parte del risultato di questa operazione è riportato in Appendice C . Vediamo che nella matrice compare anche lo stato -1 non definito precedentemente, esso viene generato da situazioni particolari. Innanzitutto rappresenta lo stato iniziale

per quelli di cui non è possibile la conoscenza di quello reale ovvero per gli episodi censurati a destra generati dagli episodi in corso al gennaio 1994 ma anche da quelli generati da individui che entrano nel panel in tempi successivi. Questo spiega anche perché ci sono individui che hanno il momento iniziale del primo episodio osservato in un tempo successivo a  $t = 0$ . All'opposto ci permette anche di definire gli episodi censurati a sinistra generati dalle uscite dal panel. In sostanza lo stato indefinito  $-1$  identifica gli episodi censurati non dalla finestra di osservazione ma dalle uscite e dagli ingressi nel panel interni a questa finestra. In ogni caso, una volta identificato il tipo di censura, destra o sinistra o entrambi, questi vengono considerati al pari di quelli censurati dalla finestra di osservazione.

Questa matrice contiene 19685 episodi di tutti i tipi, anche episodi che contano i mesi in cui l'individuo è stato nello stato  $-1$  ovvero non è stato intervistato per un periodo per poi rientrare a far parte del campione. Eliminando questi episodi si ottengono 16262 episodi rappresentanti 6 diversi tipi di transizione. Selezionando solo gli episodi che identificano una transizione dallo stato 2=disoccupato a quello 1=occupato oppure che partono dallo stato 2 e sono censurati si ottengono 4243 episodi di disoccupazione competiti o incompleti.

Il calendario delle attività relativo al 1993 verrà utilizzato per recuperare i momenti di inizio di episodi che risultano censurati a sinistra se osservati dal gennaio 1994 nonché per costruire parte della storia del processo lavorativo rendendo possibile ad esempio la conoscenza della durata di eventuale episodio lavorativo vissuto in questo anno.

Successiva alla costruzione della matrice di transizione è quella della matrice delle covariate individuali. Queste vengono costruite a partire dalle informazioni del *Personal file* ma anche sulla base di informazioni sulla famiglia contenute nell'*Household file*: queste verranno linkate agli episodi corrispondenti attraverso il codice identificativo dell'individuo.

### 3.3 Costruzione delle covariate

In questo paragrafo verrà esemplificato il procedimento che ha portato alla identificazione e costruzione di tutte le covariate poi utilizzate in fase di stima.

Prima di affrontare questo, però, mi sembra necessario discutere quella che è la relazione temporale tra le covariate e la transizione d'interesse.

Il tempo rappresenta uno degli elementi concettuali fondamentali di un'indagine di tipo panel: si hanno le caratteristiche degli individui ma anche quando queste sono state acquisite o modificate e quindi, in un'ottica causale, capire

se queste modificazioni portano alla modificazione anche del tasso di transizione. Ciò implica che dare o non dare la possibilità alle variabili di modificare il loro valore in corso di stima può portare a distorsioni e in ogni caso a riferire gli eventi a caratteristiche diverse da quelle che le hanno realmente influenzate. Avendo definito univocamente l'intervallo temporale di riferimento resta automaticamente definito l'intervallo che identifica il passato del processo e che andrà a costituire parte fondamentale nell'identificazione del modello. Guardando alle covariate utilizzate si possono identificare quattro *gruppi di significato*: caratteristiche individuali, caratteristiche familiari, contesto sociale e storia del processo lavorativo.

Le variabili che, indipendentemente dal gruppo di appartenenza, hanno un legame diretto con l'avanzare del tempo e che possono cambiare il loro valore man mano che il processo si evolve vengono dette *tempo-dipendenti* e richiedono una procedura apposita per la costruzione.

### 3.3.1 Costruzione delle variabili tempo-dipendenti

Per la definizione di un modello di durata è necessario definire un asse temporale sul quale osservare l'evoluzione del processo: risulta fondamentale infatti l'identificazione esatta del momento nel quale si verifica la transizione per poter identificare la durata stessa.

Molte delle variabili che si utilizzano per capire quali sono i meccanismi sottostanti determinati fenomeni sono una rappresentazione di fenomeni sociali complessi che naturalmente si evolvono. Per poter capire quali sono i riflessi di questi cambiamenti è necessario che anche la rappresentazione che di loro si fa abbia la capacità di evolversi parallelamente e in sincronia rispetto ai cambiamenti che realmente osserviamo.

Le variabili tempo dipendenti sono così dette appunto perché, per costruzione, il loro valore viene modificato in corrispondenza di determinati istanti temporali. Tali istanti vengono identificati attraverso i momenti (le date) in cui si verificano i cambiamenti del processo parallelo a quello in studio del quale si vuole verificare l'effetto.

Consideriamo un episodio di disoccupazione che inizia nel settembre del 1994 e termina in aprile 1995, quando l'individuo trova lavoro.

Vogliamo valutare l'effetto che una ricerca più o meno attiva può aver avuto sul fatto che l'individuo abbia trovato lavoro. Immaginiamo che nel 1994 l'individuo abbia attuato solo 2 delle 4 strategie a disposizione mentre nel 1995 decida di essere più attivo e attuarle tutte indiscriminatamente. L'utilizzo di una variabile che in corrispondenza di gennaio 1995 cambia valore passando da 2 a 4 permette di valutare se questo cambiamento si riflette anche in una



modificazione significativa della probabilità di transizione.

Questo concetto si attua nel metodo dell' *episode-splitting* [Blossfeld, Rohwer, 1995] che consiste nel dividere l'episodio in più pezzi in corrispondenza dei momenti nel tempo, interni alla durata osservata, nei quali si verificano eventi legati all'effetto che si vuole analizzare. Nell'esempio appena proposto il metodo si risolve nello spezzare l'episodio originario in due parti, una che va da settembre a dicembre 1994 e l'altra che va da gennaio a aprile 1995. L'episodio originario resta comunque identificato creando una sorta di struttura gerarchica che permetta di identificare, in ordine temporale, tutti i pezzi appartenenti allo stesso episodio. Praticamente si ottiene una matrice dati tanto più grande di quella originaria quanti più effetti si vogliono trattare come tempo dipendenti con delle variabile aggiuntive che permettano di mantenere le relazioni tra gli episodi originari e quelli splittati.

Per la costruzione della variabile tempo-dipendente si utilizzerà una sorta di contatore temporale interno che scandisce il tempo del processo da un'origine fissata fino alla durata massima osservata. La variabile viene indicizzata attraverso questo contatore e cambia il suo valore quando il momento scandito dal contatore corrisponde a quello in cui abbiamo effettuato lo split: è proprio qui infatti che la situazione è cambiata e la variabile deve cogliere proprio questo cambiamento.

In fase di stima questo tipo di covariate permette di valutare la variazione indotta nel rischio di periodi in cui si possedeva o meno una determinata caratteristica ovvero di valutare se il periodo di ricerca attiva ha una probabilità maggiore di sfociare in un lavoro rispetto a quello in cui la ricerca è stata meno attiva.

Questo metodo prevede in realtà due specificazione, una con split variabili e una con split fissi; i due tipi possono essere utilizzati congiuntamente e vengono definiti a seconda che tutti gli episodi vengano splittati a date fisse che vengano splittati a date dipendenti da un altro fenomeno e quindi in momenti diversi se appartenenti a individui diversi.

In questo lavoro il tempo è misurato in mesi ma le altre variabili costruite hanno come riferimento temporale per la conoscenza degli eventuali cambiamenti l'anno che intercorre tra le 2 interviste. In conseguenza di ciò gli episodi sono stati splittati con cadenza annuale ovvero ogni 12 mesi nella misura della durata. La sola variabile che non ha osservato lo stesso procedimento è quella relativa al vivere o meno in coppia poiché è stato possibile ricostruire le date, mese e anno, di inizio e fine della convivenza. in questo caso lo split è variabile e avviene per ogni episodio in corrispondenza del mese e dell'anno in cui la caratteristica varia.

### 3.4 Caratteristiche personali

Di questo blocco di variabili fanno parte quelle covariate che permettono l'identificazione degli individui attraverso le loro caratteristiche demografiche di base nonché qualche indicatore della situazione economica nel periodo.

Vorrei precisare subito che tra le variabili esplicative non sarà presente nessun indicatore della situazione degli individui rispetto all'istruzione. Ciò è dovuto al fatto che non è stato possibile ricostruire il percorso poiché nel corso delle wave è stata modificata la modalità con la quale queste informazioni venivano richieste e solo nel 1998 viene esplicitamente chiesto qual è l'ultimo titolo di studio acquisito. Questa variabile non è comunque risultata essere informativa rispetto al tasso di transizione quando inserita nel modello finale (appendice E, E.5)].

Nel gruppo che definisce le caratteristiche individuali rientrano le seguenti variabili:

- **ETÀ** : rappresenta l'età in anni compiuti al momento dell'intervista. Trattandosi di una variabile tempo-dipendente è stata costruita utilizzando il metodo dell'episode-splitting (3.3.1): a partire dal valore osservato alla prima intervista, sia essa coincidente con l'inizio della finestra di osservazione oppure no, il valore iniziale viene incrementato di 1 anno ogni dodici mesi di avanzamento del processo. È costruita a partire dalla variabile PD003 <sup>1</sup> previo controllo di congruenza con PD001, anno di nascita.
- **COORTE** di appartenenza: definita attraverso classi di anni di nascita. Le classi sono definite come COO1 se  $PD001 < 1966$ , COO2 se  $1966 \leq PD001 < 1973$  e COO3 se  $PD001 \geq 1974$  con limiti che garantivano numerosità comparabili tra i tre gruppi.
- **SEX** : codifica del sesso dell'intervistato. 0=M , 1=F.
- **COPPIA**: dummy che indica se l'individuo vive in coppia oppure no. Per la costruzione di questa variabile sono stati monitorati i valori assunti dalle variabili PD005, PD006 e PD007 in modo da ricostruire le date di inizio e fine di tutte le eventuali convivenze e quindi le traiettorie individuali rispetto a questo processo parallelo a quello in studio. A partire dalla situazione rilevata alla prima intervista la variabile cambia

---

<sup>1</sup>I codici si riferiscono ai nomi originali delle variabili nell'UDB, vedi Appendice A

valore in corrispondenza del mese e anno in cui la convivenza inizia, passaggio da 0 a 1, oppure termina, passaggio da 1 a 0.

Per come è costruita permette di dire se il tasso di transizione di un individuo varia nei periodi in cui è in coppia. La categoria di riferimento è infatti lo 0 ovvero non vivere in coppia.

- AFFITTO : indica se l'individuo deve pagare oppure no un affitto per la sua abitazione. Questa informazione viene dedotta dal file che rileva le caratteristiche famigliari: è infatti su questo che si trova una sezione dedicata alla posizione della famiglia rispetto all'abitazione. La variabile è tempo-dipendente e costruita su split fissi annuali; tra gli individui che non pagano un affitto rientrano anche i giovani che abitano nella casa paterna oltre a quelli che hanno un'abitazione di proprietà e rappresentano la classe di riferimento per la lettura delle stime.
- REGION : ripartizione geografica di residenza dell'individuo al momento dell'intervista. La disaggregazione utilizzata prevede 4 aree: Nord, Centro, Sud e Isole ed è costruita a partire da informazioni presenti nel file delle famiglie. Dati i pochi trasferimenti che davano origine, nei 5 anni di indagine, a cambiamenti di ripartizione, la variabile è stata utilizzata tempo costante con valore fisso a quello rilevato nella prima intervista.
- REDIND : reddito annuale dell'individuo .  
È ricostruito a partire dal controllo sulle variabili della sezione *INCOME* dei *personal file* e prevede quattro classi costruite a partire dai quartili della distribuzione osservata. Anche questa variabile è tempo dipendente in quanto il reddito annuale può risultare diverso in anni diversi. Presumibilmente il reddito aumenterà per quegli anni in cui si hanno degli episodi di lavoro.

### 3.5 Caratteristiche famigliari

Le variabili contenute in questo gruppo sono state ricostruite a partire dalle informazioni contenute nell'*Household file* e sono un tentativo di rappresentare alcune situazioni della famiglia dell'individuo intendendo con famiglia quella nella quale vive al momento dell'intervista sia essa quella d'origine o quella costruita a seguito di una unione.

- BIMBI : numero di bambini con meno di 12 anni facenti parti della famiglia. L'informazione è stata trattata come tempo dipendente ed è stata inizialmente associata ad un'altra variabile che coglieva la presenza di individui con meno di 15 anni. Tra le due la prima è risultata essere maggiormente discriminante rispetto ai livelli di rischio.
- REDFAM : reddito familiare annuale. È costruito sulla base dei quartili della distribuzione osservata del reddito familiare; è stata costruita sulla scia di quello individuale a partire dalla corrispondente sezione *IN-COME* dell'*Household file*, ed è anch'essa tempo-dipendente. Per alcuni individui che formano famiglia a se è coincidente con il reddito individuale.
- SIZE: numero di componenti della famiglia. Sono state costruite 6 classi di cui l'ultima aperta raggruppante gli individui con famiglie di 6 o più componenti. La variabile è stata costruita tempo dipendente con l'intento di eliminare le eventuali distorsioni provocate dalla modificazione delle strutture famigliari degli intervistati.

### 3.6 Contesto sociale

Sotto questo gruppo si trovano alcune variabili indicatrici della situazione economica dell'Italia; tutte sono state costruite tempo-dipendenti per riflettere il cambiamento del loro valore passando da un anno al successivo.

- RATEGIOV : tasso di disoccupazione giovanile dell'Italia.
- RATETOT : tasso di disoccupazione totale dell'Italia.
- PIL : prodotto interno lordo pro-capite per l'Italia durante gli anni dell'osservazione.

### 3.7 Il passato del processo

Come è stato illustrato nel par. 2.3.3 il modello che si va ad utilizzare vorrebbe rappresentare una generalizzazione dei processi semi-markoviani i quali

considerano ininfluenza rispetto all'evoluzione futura la storia del processo precedente il momento in cui viene attraversato lo stato al quale siamo interessati.

A livello empirico ciò si riflette nell'inserimento di covariate rappresentanti alcuni aspetti di questo passato; è difficile valutare a priori quali sono le informazioni importanti quindi si procederà considerando da un lato le opportunità offerte dai dati a disposizione e dall'altra la validazione che di esse si può ottenere in fase di stima. Potrebbe infatti verificarsi che nessuna delle variabili riguardanti questo aspetto del processo influenzi la probabilità di transizione e che quindi il fenomeno sia perfettamente rappresentabile attraverso un processo semi-markoviano. D'altro canto potrebbe essere che le variabili rappresentanti il passato del processo siano influenti in quanto non in grado di cogliere gli aspetti importanti di esso ai fini della determinazione del tasso di transizione.

- SERIAL : numero progressivo di osservazione dell'episodio; prende valori da 1 a 4 dove quest'ultimo rappresenta la classe aperta degli individui che vivono 4 o più episodi. La costruzione di questa variabile discende dalla costruzione della matrice di transizione e dalla numerazione progressiva degli episodi per ogni individuo.
- DURPREJ : durata dell'episodio di occupazione precedente quello di disoccupazione che si sta analizzando. Viene costruita sulla matrice di transizione e per i primi episodi l'episodio di occupazione eventuale viene recuperato dal calendario relativo al 1993 non utilizzato in fase di costruzione della matrice di transizione.  
La durata dell'occupazione è stata costruita tempo-dipendente e vale zero se non è stato osservato nessun episodio di occupazione precedente quello di disoccupazione in analisi. Se il secondo episodio di disoccupazione non è preceduto da un periodo di occupazione ma ad esempio di inattività la variabile torna a zero.
- DURPREU : durata dell'episodio di disoccupazione precedente quello in analisi. Anche questa variabile è stata costruita sulla base della matrice di transizione: vale zero in corrispondenza del primo episodio osservato e prende il valore della durata dell'episodio  $i$  in corrispondenza dell'episodio  $i + 1$ .

- **ACTIVE** : numero di azioni svolte durante l'anno nella ricerca di un'occupazione. Dalla sezione *SEARCH FOR WORK* del questionario individuale sono stati identificati 4 possibili *azioni/atteggiamenti* possibili nei confronti della ricerca del lavoro e la variabile indica quanti di questi sono stati utilizzati. La variabile è tempo-dipendente e cambia il suo valore in corrispondenza degli anni nei quali si osserva un cambiamento nelle variabili utilizzate per la loro costruzione; vale zero se non è stata attuata nessuna delle strategie proposte per cui potrà informarci rispetto alle modificazioni del rischio per chi attua una o due etc... delle possibilità identificate dal questionario.
- **UNEM5P** : dummy indicante se l'individuo ha avuto o meno qualche periodo di disoccupazione nei 5 anni precedenti la prima intervista costruita sulla base di PU002, variabile della sezione *UNEMPLOYMENT* del questionario personale (Appendice A).  
Resta fissa al valore iniziale per tutto il periodo di osservazione in quanto mentre le tre variabili precedenti colgono l'evoluzione del processo nel corso dell'osservazione e quindi costituiscono un passato che si modifica e si amplia all'avanzare dell'osservazione, questa, come le successive rappresentano delle informazioni statiche relative ad un periodo, i 5 anni precedenti l'intervista del quale non abbiamo nessuna osservazione diretta.
- **NUMUNEM** : numero di volte che l'individuo è stato disoccupato nei 5 anni precedenti l'intervista. Scala da 1 a 6: 6 e più episodi. Rimane fissa al valore iniziale costruito sulla base di PU003. A questa variabile è stata inizialmente associata anche un'altra variabile che, a partire dal valore di NUMUNEM si incrementava di uno in corrispondenza di ogni episodio di disoccupazione osservato. Questa variabile se inserita da sola dava lo stesso effetto, anche se meno forte, della prima con la quale risultava collineare se inserita in un modello nel quale NUMUNEM era già presente. Si è deciso di tenere quindi la prima anche per la coerenza che ha, a livello temporale con UNEM5P e LONYEAR.
- **LONYEAR** : dummy indicante se uno degli episodi di disoccupazione vissuti nei 5 anni precedenti la prima intervista è stato più lungo di un anno (PU004).

- NUMLONYEAR : numero di episodi osservati di lunghezza superiore ad un anno. Al valore iniziale costruito sulla base della matrice di transizione si aggiunge un episodio se l'individuo dichiara di averne avuto almeno uno nei 5 anni precedenti l'ingresso nel panel; la sua tempo dipendenza è legata alla cardinalità dell'episodio nel senso che si modifica in corrispondenza di un episodio se il precedente è stato di durata maggiore o uguale a 12 mesi.





# Capitolo 4

## Analisi empirica e risultati

### 4.1 Descrizione degli episodi

La popolazione di riferimento per questa analisi è data dagli individui del campione che hanno meno di 35 anni alla prima intervista sia che avvenga alla prima ondata che in una ondata successiva. La selezione di questi individui porta a lavorare con un insieme di 2359 persone <sup>1</sup> che genera i 4243 episodi di disoccupazione che andremo ad analizzare.

La prima esplorazione dei dati è stata condotta attraverso la definizione di un sistema ad episodi multipli. Questa struttura permette infatti di non perdere la relazione temporale tra gli episodi successivi di uno stesso individuo nel senso che, definito l'asse temporale, gli episodi vengono posizionati esattamente nel punto in cui hanno origine nella realtà: quando da questa struttura reale a multi episodio si passerà, per i motivi discussi nel par. 2.2, ad una struttura ad episodi singoli, la relazione temporale tra gli episodi generati da uno stesso individuo si perde in quanto viene fissata un'origine comune per tutti.

Dai dati risulta che la cardinalità massima osservata è di 13 episodi, vissuti però da due soli individui. La numerosità dei gruppi definiti attraverso la cardinalità è funzione decrescente della stessa e passa da 2359 per i primi episodi a 95 per quelli di cardinalità quattro per restare poi sempre sotto i 95 episodi. gli episodi sono stati quindi raggruppati utilizzando solo 4 classi di cardinalità con l'ultima che racchiude tutti gli episodi di cardinalità osservata maggiore o uguale 4.

La distribuzione risultante è

---

<sup>1</sup>Questo è l'insieme di coloro che hanno meno di 35 anni ma che hanno anche informazioni complete per tutte le variabili che poi verranno utilizzate come covariate. In realtà senza il vincolo della completezza delle informazioni nel corso degli anni l'insieme sarebbe stato di 6174 individui al 1994 per ridursi a 4180 nel 1998

CARDINALITÀ	ORG	DEST	N EPISODI	DURATA MEDIA
1	0	0	1009	26.87
1	0	1	1350	11.89
SUM			2359	
2	0	0	314	13.04
2	0	1	452	7.01
SUM			766	
3	0	0	95	9.21
3	0	1	152	4.83
SUM			247	
4	0	0	57	4.16
4	0	1	120	4.37
SUM			177	

Tabella 4.1: Distribuzione degli episodi per cardinalità osservata. Lo stato 0 indica che l'individuo è disoccupato e lo stato 1 che è occupato

Accanto alla distribuzione seriale degli episodi sono indicate anche le durate medie sia per coloro che infine trovano lavoro sia per coloro che restano disoccupati fino alla fine del periodo di osservazione. Anche se la media è un indicatore non robusto possiamo osservare che la durata media del periodo di disoccupazione decresce all'aumentare della cardinalità dell'episodio: più episodi si vivono e più questi sono brevi ma non possiamo dire nulla sulla probabilità di transizione o sulla velocità di eliminazione del campione. Per rispondere a queste domande ci si avvale della stima non-parametrica delle funzioni di sopravvivenza, dalle quali ricavare anche una stima del rischio, attraverso il metodo di Kaplan-Meyer o del prodotto limite. L'idea base di questo metodo è di considerare tutti gli istanti di tempo in cui si verifica almeno un evento, diciamo  $q$  punti,  $\tau_1 < \tau_2 < \dots < \tau_q$  con la convenzione che  $\tau_0 = 0$  e definire  $\forall j \in (1, 2, \dots, q)$  le quantità:

- $E_j$ : numero di episodi che terminano con un evento in  $\tau_j$
- $Z_j$ : numero di episodi censurati che terminano nell'intervallo  $[\tau_{j-1}, \tau_j]$
- $R_j$ : risk set in  $\tau_j$  ovvero tutti gli episodi che iniziano in un istante inferiore a  $\tau_j$  e terminano in un istante di tempo maggiore o uguale a  $\tau_j$ <sup>2</sup>.

<sup>2</sup>In questo risk set sono inclusi anche gli episodi censurati in  $\tau_j$ : l'assunzione è che anche questi sono informativi in quanto ci dicono che non ci sono eventi fino a *e incluso* il tempo osservato di fine dell'episodio.

Date queste quantità, lo stimatore della funzione di sopravvivenza è definito come

$$\hat{S}(t) = \prod_{j:\tau_j < t} \left(1 - \frac{E_j}{R_j}\right)$$

risultante in una funzione a gradini con salti in ogni istante di tempo  $\tau_j$  nel quale si è verificato almeno un evento.

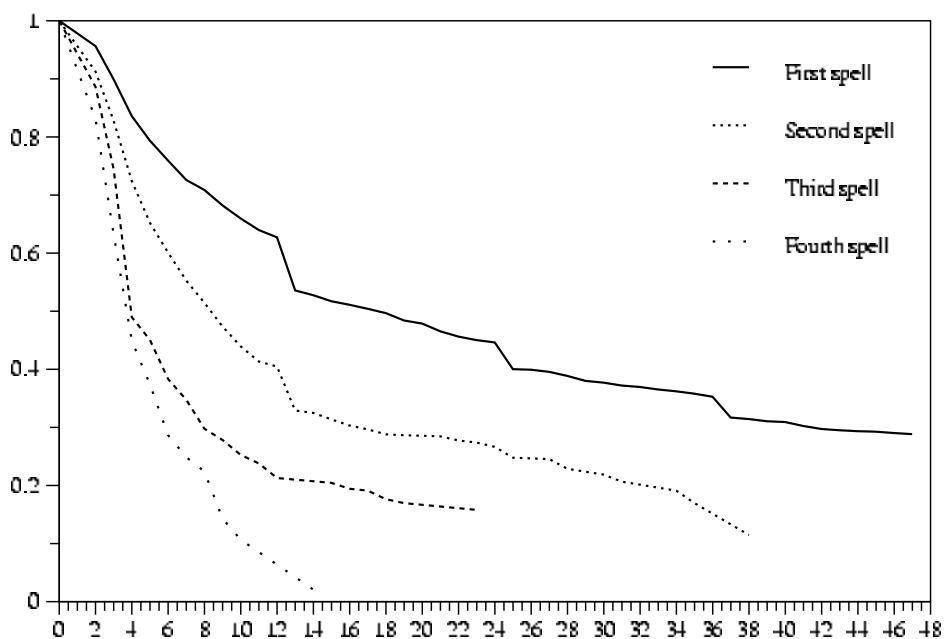


Figura 4.1: Funzione di sopravvivenza per cardinalità dell'episodio

Dal calcolo della funzione di sopravvivenza possiamo ottenere la mediana della distribuzione e quindi vedere se i sottogruppi definiti attraverso la cardinalità dell'episodio hanno una velocità di eliminazione diversa. Le mediane risultanti si scostano abbastanza dalle medie osservate: per il primo episodio abbiamo che a fronte di una durata media del periodo di disoccupazione di circa un anno la transizione avviene per metà del campione non prima i 17 mesi e mezzo dall'inizio del periodo di disoccupazione.

Riassumendo abbiamo i seguenti valori:

Se fissiamo un punto nel tempo possiamo osservare che la funzione di sopravvivenza condizionata alla cardinalità dell'episodio, diciamo  $\hat{S}(t|i)$  con

CARDINALITÀ	DURATA MEDIANA	DURATA MEDIA
1	11.89	17.55
2	13.04	8.36
3	9.21	3.96
4	4.16	3.57

Tabella 4.2: Durate mediane e medie per cardinalità dell'episodio

$i \in 1, 2, 3, 4$ , è funzione decrescente del tempo e che non c'è mai incrocio tra le curve di episodi diversi. Gli episodi che succedono il primo sono più corti e mostrano un tasso di transizione maggiore.

Per vedere se le distanze tra le curve rappresentano una differenza significativa di comportamento è necessario confrontarle attraverso dei test. Se le differenze non sono significative allora è possibile lavorare su un insieme unico e non stratificato attraverso la cardinalità.

I test esposti nel par.2.4.1 hanno dato i seguenti risultati:

test statistic	T-stat	DF	$\alpha$
Log-Rank (Savage)	390.6480	3	0.0000
Wilcoxon (Breslow)	371.2377	3	0.0000
Wilcoxon (Tarone-Ware)	386.4742	3	0.0000
Wilcoxon (Prentice)	378.7567	3	0.0000

Tabella 4.3: Test statistici per il confronto delle funzioni di sopravvivenza condizionate alla cardinalità osservata dell'episodio

I risultati mostrano l'esistenza di una differenza significativa tra le curve quando queste vengono considerate nell'insieme, ma per vedere dove effettivamente si concentra la differenza, i test sono stati calcolati per coppie di funzioni di sopravvivenza e i risultati mostrano che l'entità delle differenze non è sempre la stessa: pur risultando sempre significativi, la differenza acquista maggiore significatività quanto più gli episodi sono distanti tra loro. L'osservazione delle curve di sopravvivenza condizionate alla cardinalità dell'episodio ci mostra che la legge di eliminazione non è la stessa e che, di conseguenza, anche le rispettive funzioni di rischio mostreranno livelli differenti in corrispondenza degli stessi valori di durate nonchè, probabilmente, forme differenti.

Le funzioni di sopravvivenza condizionate mostrano delle differenze di comportamento anche rispetto ad altre variabili che rappresentano diversi aspetti della storia del processo.

Se si restringe l'attenzione ai 5 anni precedenti l'intervista, si possono considerare separatamente coloro che in questo periodo sono stati disoccupati una

o più volte, da coloro che invece si trovano in questa condizione per la prima volta. In Appendice A fig.D.1 sono riportate le curve stimate e il valore dei test (tab. D.1) che mostrano l'esistenza di una differenza significativa. Coloro che si trovano al primo episodio di disoccupazione degli ultimi 5 anni mostrano una transizione più veloce<sup>3</sup> e di intensità finale maggiore rispetto a coloro che invece, nello stesso periodo, si sono trovati a fronteggiare più volte questa situazione. Queste stime sembrano indicare l'esistenza di un effetto del passato sulla futura propensione del processo a cambiare stato.

Altre considerazioni interessanti riguardano l'effetto che un periodo di occupazione ha sul successivo periodo di disoccupazione, fig. D.2 . I gruppi sono identificati dalla presenza o meno tra primo e secondo episodio di un periodo di occupazione, la base per la stima è quindi costituita dagli episodi successivi al primo. Anche in questo caso le funzioni sono significativamente diverse sia nei livelli che raggiungono sia nella velocità con la quale lo fanno; test in tab. D.2.

Coloro che fanno registrare un episodio di lavoro tra i due primi episodi osservati di disoccupazione mostrano una propensione maggiore a trasformare il secondo episodio, come hanno fatto con il primo, in un lavoro. Le durate mediane sono nettamente differenti e passano da 7,02 mesi per il primo gruppo a ben 33 per il secondo. Probabilmente questo secondo gruppo è eccessivamente eterogeneo contenendo da coloro che tra i due episodi di disoccupazione hanno seguito un corso di formazione a coloro che hanno prodotto 2 episodi in quanto scoraggiati dall'esito negativo della prima ricerca sono in qualche modo usciti dal mercato per poi rientrarvi di nuovo come disoccupati.

Lo stesso risultato si ottiene se si considera la probabilità di transizione al lavoro dal terzo episodio osservato di disoccupazione per coloro che hanno lavorato o meno nel periodo intercorso tra il secondo e il terzo. Anche queste curve sembrano dimostrare che la probabilità di uscire da un periodo di disoccupazione non è indipendente da ciò che è successo nel periodo subito precedente.

Diamo uno sguardo infine alle funzioni di sopravvivenza di un episodio condizionate alla durata del precedente. In fig. D.3 sono riportate quelle relative ai secondi episodi distinguendoli a seconda che il primo sia stato più o meno lungo di un anno. Se il primo episodio è stato più corto di un anno la probabilità di transizione al lavoro per il secondo episodio di disoccupazione è maggiore in corrispondenza di qualsiasi durata ed avviene per metà del sotto-campione entro i primi 8,7 mesi a differenza dell'altro gruppo nel quale bisogna aspettare ben 2 anni prima che la stessa percentuale abbia trovato

---

<sup>3</sup>per questo gruppo la durata mediana è circa la metà di quella dell'altro gruppo: 5,64 mesi contro 12,86

lavoro. Anche il livello finale della curva mostra che aver avuto un primo episodio di disoccupazione non più lungo di un anno è un vantaggio.

Da queste stime sembra quindi che il passato del processo, almeno negli aspetti qui considerati, non è completamente ininfluenza sull'attuale probabilità di transizione; sarà ovviamente la stima del modello che, potendo cogliere anche gli effetti combinati delle variabili, sarà in grado di dire se e in che verso questo passato influenza l'evoluzione futura del processo.

## 4.2 Scelta della forma parametrica

La scelta della forma parametrica da utilizzare per rappresentare la funzione di rischio si basa su una sua preliminare stima attraverso un metodo non parametrico. Tale stima viene derivata da quella della funzione di sopravvivenza passando attraverso una stima del rischio cumulato che si otterrà come  $H_T(t) = -\ln S_T(t)$ , da questa, derivando in funzione del tempo si ottiene una stima del funzione di rischio poiché  $H_T(t) = \int_0^t h_t(u)du$  e quindi  $h_t(t) = \frac{dH_T(t)}{dt}$ .

La funzione di sopravvivenza è una funzione a gradini e tale risulterà anche la funzione di rischio cumulato che quindi non è derivabile in alcuni punti, esattamente quelli in cui si sono verificate delle transizioni. Per ovviare a questo problema la funzione di rischio cumulato è stata resa continua attraverso l'applicazione di una funzione di smooth, nello specifico una spline cubica, e poi è stato applicato l'algoritmo di derivazione; la funzione risultante è presentata in fig. 4.2 .

La funzione risultante è bimodale ; per la prima parte ripercorre la forma tipica della funzione di transizione all'occupazione ovvero con una prima parte crescente fino ad un massimo per poi ridiscendere. Nell'ultima parte della curva si osserva una seconda gobba che potrebbe evidenziare un particolare legato alla disoccupazione di lungo periodo ma potrebbe anche essere dovuto ad un concentramento casuale di transizioni che pesa molto poiché gli episodi ancora in corso e quindi a rischio sono pochi.

Il primo passo è stato quello di utilizzare i test grafici presentati nel par.2.4.1 utilizzando più funzioni con l'idea di riuscire già in questa fase ad avere indicazioni almeno di quelle palesemente non adatte.

Le funzioni prese in considerazione in questa prima fase della scelta sono state: Weibull, Log-logistica, Log-normale.

Nei test grafici si utilizza una trasformata della funzione di sopravvivenza che

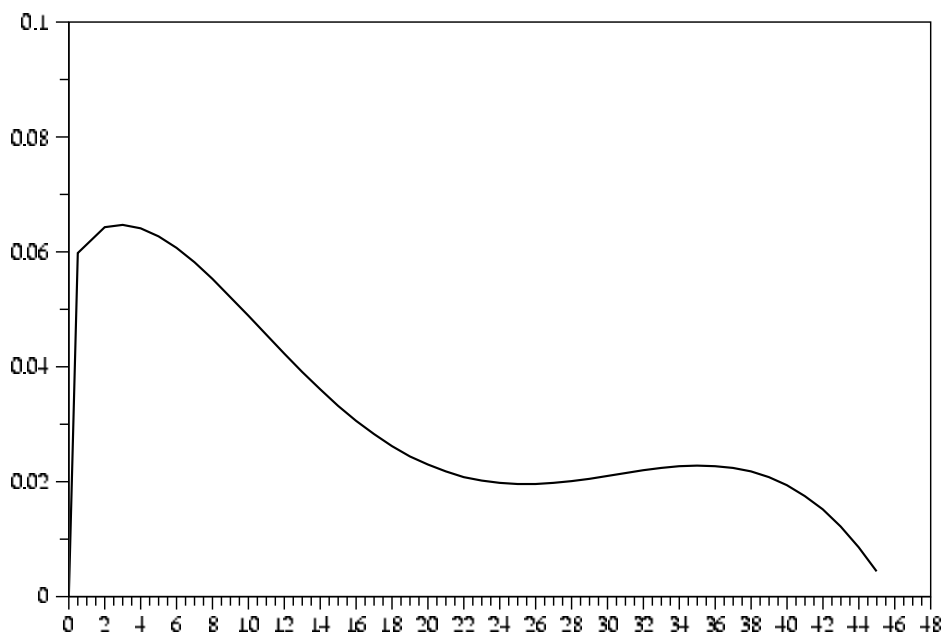


Figura 4.2: Stima della funzione di rischio per l'insieme degli episodi

sia in relazione lineare con il tempo o una trasformata dello stesso. L'adattamento viene valutato da un punto di vista grafico ma anche con l'ausilio dell'indice di determinazione lineare calcolato dalla regressione che prende la trasformata della sopravvivenza come variabile dipendente (fig. 4.3).

L'adattamento migliore sembra essere quello della funzione log-normale che differenzia il suo comportamento soprattutto nel tratto iniziale della curva. L'indice di determinazione lineare pari a 0.98 ed è il più alto dei 3. Il passo successivo è stato quello di stimare dei modelli marginali, ovvero senza covariate, utilizzando una specificazione parametrica del rischio di base. Con questa procedura si stimano, a partire dai dati, i parametri della funzione utilizzata e si può quindi disegnarla: i risultati per le distribuzioni log-logistica e log-normale sono presentati in fig.4.4 mentre i valori delle stime dei parametri in tab. 4.4<sup>4</sup>.

---

<sup>4</sup>La distribuzione weibull è stata scartata essendo quella che dimostra il minor adattamento. Ciò era prevedibile poichè la funzione di rischio stimata non-parametricamente risulta essere campanulare mentre la distribuzione weibull determina una funzione di rischio monotona.

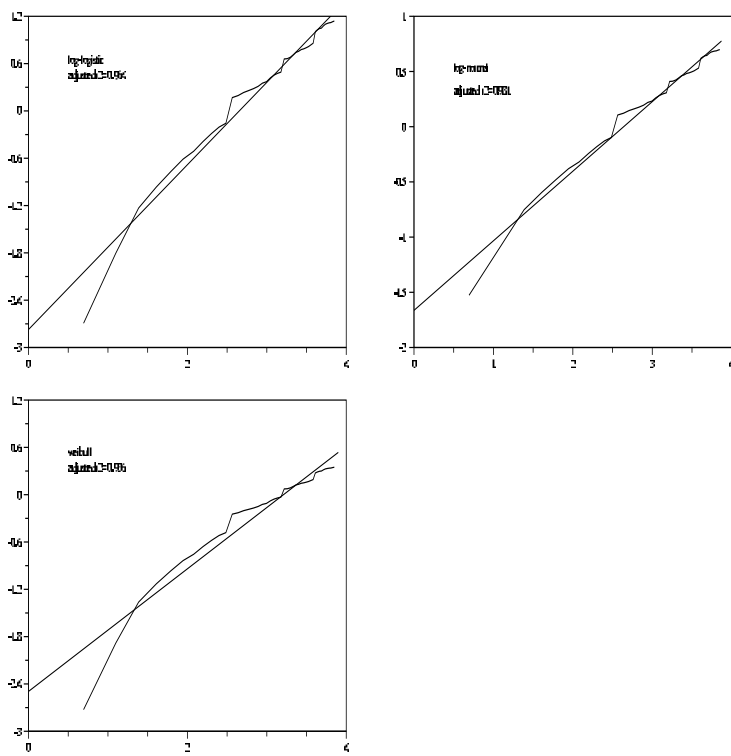


Figura 4.3: Confronto dei test grafici per le funzioni log-logistica, log-normale e weibull.



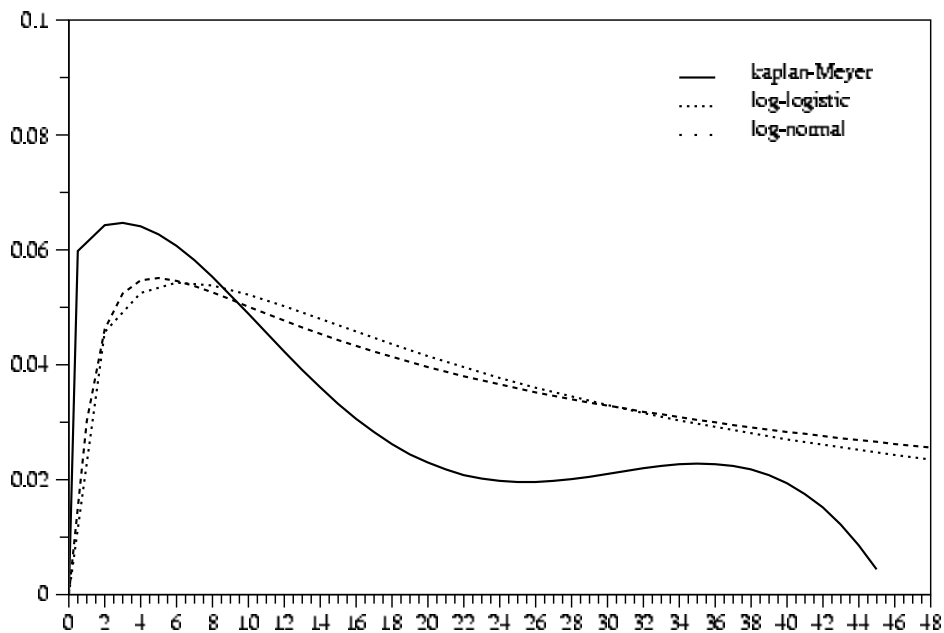


Figura 4.4: Funzione di rischio stimata con diversi modelli di riferimento: log-normale, log-logistico e non-parametrico

Dai modelli così stimati sono stati calcolati gli pseudo-residui nella formulazione data in par.2.4.2; graficamente valutiamo se la funzione di rischio risultante dalla distribuzione degli pseudo-residui è costante ovvero se  $-\ln(\hat{S}_e(\hat{e}))$  vs  $\hat{e}$  approssima una retta (fig. 4.5).

I risultati mostrano che i residui meno informativi si hanno con la definizione log-normale del rischio di base, meglio che con una definizione log-logistica la distribuzione dei residui approssima quella esponenziale. L'indice di determinazione lineare della regressione tra la trasformata della funzione di sopravvivenza dei residui e i residui stessi è pari a 0.9606 per la definizione log-normale e a 0.9570 per quella log-logistica.

I risultati presentati mostrano che una definizione log-normale come distribuzione della variabile di durata permette un buon adattamento alla funzione di rischio del campione; il modello che verrà utilizzato assumerà questa forma parametrica per la definizione del rischio di base e ci permetterà di vedere come media e varianza di base vengono modificati dalla presenza delle covariate.

log-normale	$h_T(t) = \frac{1}{\sigma t} \frac{\varphi(\frac{\ln t - \mu}{\sigma})}{1 - \Phi(\frac{\ln t - \mu}{\sigma})} =$	
parametro	stima	standard error
$\mu$	2.6780	0.0246
$\sigma$	1.2242	0.0172
log-likelihood=-7756.4545		
log-logistica	$h_T(t) = \frac{\beta \alpha t^{\beta-1}}{1 + (\alpha t)^\beta}$	
$\alpha$	0.0715	0.0251
$\beta$	1.3450	0.0187
log-likelihood=-7850.8709		

Tabella 4.4: Stima dei parametri della funzione di rischio di base nella definizione log-normale e log-logistica. Tutte le stime sono significativi all' 1 per mille

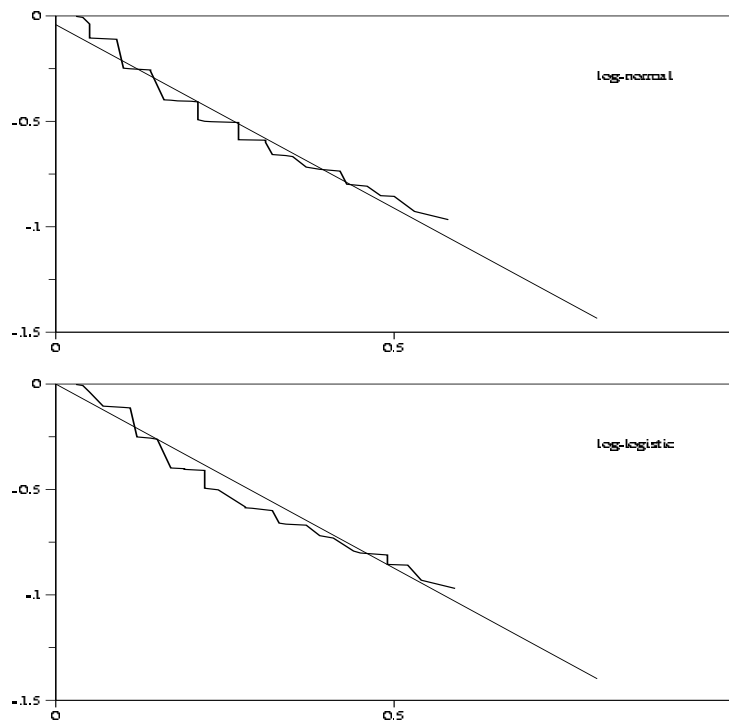


Figura 4.5: Test sugli pseudo-residui per le forme parametriche log-normale e log-logistica

### 4.3 Definizione e stima del modello

Il modello utilizzato è definito attraverso una distribuzione log-normale per il rischio di base e un termine di errore distribuito come una variabile gamma. Questo tipo di modello permette di legare gli effetti delle covariate sia alla media che alla varianza della distribuzione risultante per il rischio quindi i parametri stimati rappresenteranno l'entità e il verso della variazione indotta dalle covariate sui parametri del rischio di base.

Nella formulazione utilizzata il modello è descritto come

$$h_T(t) = h_T^b(t|x) \left[ 1 + \frac{1}{k} H_T^b(t|x) \right]^{-1}$$

dove  $h_T^b(t|x) = \frac{1}{\sigma t} \frac{\varphi(z_t)}{1 - \Phi(z_t)}$  e  $z_t = \frac{\ln t - \mu}{\sigma}$

$\varphi$  e  $\Phi$  rappresentano rispettivamente densità e distribuzione di una normale standardizzata mentre  $k$  è il parametro derivante dalla distribuzione gamma utilizzata per controllare il termine di eterogeneità (par. 2.5.1). Il legame tra le covariate e i parametri della log-normale avviene in forma esponenziale ovvero ponendo

$$\mu = \mu(x) = \exp(x\alpha) \quad \sigma = \sigma(x) = \exp(x\beta)$$

Nel modello utilizzato si legheranno le covariate al parametro  $\mu$ . Una volta definite le forme parametriche che costituiranno la distribuzione di mistura del rischio di base e l'insieme delle covariate, si procederà alla stima dei parametri del modello attraverso il metodo della massima verosimiglianza. Il primo passo è stato quello di stimare il modello di base, che fornisce cioè la funzione di rischio di riferimento rispetto ai parametri della quale si andranno a valutare le modificazioni prodotte dalle covariate.

variabile	(+/-)exp(coeff.)	std(coeff.)	$\alpha$
$\mu$	7.026	0.0344	0.0000
$\sigma$	0.6625	0.0343	0.0000
$k$	4.0993	0.0609	0.0000
log-likelihood=-7611.5697			

Tabella 4.5: Stima dei parametri della funzione di rischio di base. La varianza della distribuzione gamma è stimata con  $\frac{1}{k}$ . Gli  $\alpha > 0.001$  indicano variazioni non significative.

Questo modello stima in circa 7 mesi la durata media degli episodi di disoccupazione e si avvicina abbastanza alla stima che se ne ottiene a partire

dalla funzione di sopravvivenza, pari a circa 10 mesi. Se non si inserisce il termine di eterogeneità la media della durata viene sovrastimata pari a circa 14 mesi e gli errori standard delle stime mostrano valori più elevati.

In base a quanto esposto nel capitolo 2, questo modello di base andrebbe considerato valido solo condizionatamente alla cardinalità dell'episodio; questa condizione è fondamentale per poter considerare gli episodi indipendenti e modellizzarli come tali. Sulla base di questo modello, le cui stime sono presentate in Appendice E, si procederà con il testare l'effetto delle altre variabili.

Il modello (tab. E.1 Appendice E) mostra che la durata media degli episodi di cardinalità maggiore di 1 è sempre inferiore a quella del primo episodio, risultano quindi medie decrescenti come già evidenziato dalle stime delle funzioni di sopravvivenza condizionate.

Sulla base di questo modello sono state via via inserite, seguendo i blocchi individuati nel capitolo 3, tutte le altre covariate tenendo per ultime quelle relative alla storia del processo per poter identificare quanto il loro effetto abbia importanza sul tasso di rischio ma anche se la bontà di adattamento del modello ai dati migliora sensibilmente oppure no: in quest'ultimo caso non avrebbe senso appesantire il modello con questo aspetto ovvero il processo di durata della disoccupazione osservato rispetta ipotesi di tipo markoviano.

Per quanto riguarda le stime dei parametri del modello di base non si osservano valori più elevati ma non cambia il loro verso o la loro significatività (tab. E.2 Appendice E).

Un effetto forte e significativo lo si osserva per le variabili che rappresentano la ripartizione geografica di residenza. Se confrontate con il Nord Italia, tutte le altre ripartizioni mostrano una durata media più lunga del periodo di disoccupazione, la differenza è più marcata per il sud.

Mentre le variabili che indicano la coorte di nascita non sembrano avere un effetto significativo, l'effetto è visibile sulla variabile tempo-dipendente AN-NI, che rappresenta l'età esatta dell'individuo nel momento in cui ha vissuto l'episodio di disoccupazione. All'aumentare di un anno di età diminuisce di un mese la durata media dell'episodio. Un effetto indicativo è quello relativo alla variabile sesso che indica una durata media dell'episodio maggiore per le femmine che per i maschi.

Il terzo modello presentato considera l'inclusione, tra le variabili esplicative, di quelle relative ad alcuni aspetti della famiglia nella quale l'individuo vive. Queste variabili rilevano l'ampiezza della famiglia e la presenza di bambini che hanno meno di 12 anni nonché il reddito medio familiare annuale (tab. E.3).

Rispetto al precedente non si hanno grosse modifiche sui parametri delle

variabili già presenti se non una riduzione della significatività della stima del parametro  $\sigma$  ad indicare che le covariate riescono a cogliere buona parte della variabilità della durata attorno alla media. Seppur non significativa, la variabile relativa al reddito familiare indica una diminuzione della durata media all'aumentare di tale reddito. Anche l'effetto del reddito individuale sulla durata è negativo. In realtà potrebbe essere legato al fatto che redditi personali più elevati sono legati alla presenza, nel corso dell'anno, di episodi di occupazione, che, come vedremo hanno effetto di ridurre la durata media dell'episodio di disoccupazione aumentandone la probabilità di terminare con un lavoro.

Un effetto significativo si riscontra in corrispondenza delle variabili che colgono l'ampiezza della famiglia nella quale si vive e un aspetto della sua struttura attraverso la presenza di bambini al di sotto dei 12 anni. Per entrambe le variabili si ha un effetto di riduzione della durata media che potrebbe essere dovuto, ad esempio, ad un impegno maggiore nella ricerca del lavoro poiché una famiglia più o meno numerosa richiede risorse, non solo economiche, diverse.

A questo modello sono state aggiunte le variabili che in qualche modo possono cogliere l'aspetto economico del paese ovvero il tasso di disoccupazione totale e il Prodotto Interno Lordo espresso in milioni di euro, tab. E.4.

Anche in questo caso i parametri delle variabili già presenti nel modello non si modificano molto né nella direzione né nell'intensità dell'effetto colto. Gli effetti delle variabili economiche non sono molto chiari in quanto indicano una diminuzione del periodo medio di disoccupazione per i giovani a fronte di un aumento del tasso di disoccupazione totale della popolazione italiana e un aumento dello stesso all'aumentare del PIL. È da notare che la variabile ANNI perde la sua significatività e questo probabilmente perché poteva in realtà essere proprio una proxy dei cambiamenti strutturali che si verificano di anno in anno.

Nell'ultimo modello presentato si considerano tutte le variabili che possono raccogliere informazioni sull'evoluzione passata del processo di disoccupazione nonché una differenzialità in termini di impegno attivo nella ricerca di un lavoro. I risultati presentati in Appendice E tab. E.5 mostrano degli effetti interessanti e anche modificazioni sui parametri delle variabili già presenti nel modello.

esaminiamo innanzitutto l'effetto delle variabili relative al processo per poi andare a vedere quali effetti hanno avuto sulle altre covariate.

Osserviamo innanzitutto che una ricerca più o meno attiva del lavoro, ovvero che utilizza una o più strategie di quelle proposte in fase di intervista, diminuisce significativamente la durata del periodo di attesa prima di trovare un impiego. Lo stesso effetto si osserva per quanto riguarda la durata della precedente

occupazione nonché disoccupazione. In entrambi i casi si osserva che all'aumentare della durata della precedente occupazione e disoccupazione diminuisce la durata media dell'episodio in corso. Un effetto favorevole sulla propensione a trasformare l'occupazione in disoccupazione si osserva anche per gli individui che hanno già vissuto periodi di disoccupazione negli anni precedenti e questo sembra contraddire il risultato della stima delle funzioni di sopravvivenza in fig. D.1. In combinazione a questi effetti positivi si osserva un effetto negativo, nel senso che allunga il periodo medio di disoccupazione, per quegli individui che hanno un elevato numero di episodi di disoccupazione nei 5 anni precedenti l'episodio in corso. All'aumentare del numero di episodi vissuti aumenta anche la durata media dei successi. Sembra quindi che l'essere stati già disoccupati nei periodi precedenti aiuti ad uscire dalla disoccupazione attuale ma non se questi periodi sono stati numerosi: numerose entrate ed uscite dal lavoro a lungo andare non sembrano avere un effetto positivo. Per quanto riguarda l'esperienza dei 5 anni precedenti l'inizio dell'osservazione del processo non è discriminante la durata degli stessi in quanto la variabile che distingue tra la presenza o meno di episodi più lunghi di un anno non ha effetto significativo seppur l'indicazione è verso un aumento delle durate successive.

L'inserimento di queste variabili rende completamente ininfluenza l'effetto dei redditi, sia individuali che familiari nonché quello del sesso: a parità di esperienza non sembrano verificarsi discriminazioni in questo senso. Su altre variabili l'effetto è quello di invertirne il segno e quindi l'effetto sulla durata media. Ciò avviene ad esempio per le variabili di coorte che, seppur non significative, invertendo il segno del loro effetto riconoscono a mio avviso l'effetto del passato in quanto, se non lo si considera i preferiti sono sicuramente i disoccupati più giovani ma quando lo si fa, viene riconosciuta ai giovani disoccupati 'meno giovani' l'esperienza accumulata. Si inverte anche l'effetto del tasso di disoccupazione totale che indica, questa volta, un aumento della durata media della disoccupazione all'aumentare del tasso stesso.

Le variazioni più importanti si osservano però rispetto ai parametri della distribuzione del rischio di base. Si riporta alla significatività quello relativo alla media mentre resta ininfluenza quello relativo alla varianza: ancora una volta sembra che le variabili considerate riescano a cogliere la variabilità della durata attorno alla media. Una variazione di valore ma soprattutto di significatività si osserva per il parametro utilizzato per la stima della varianza della funzione gamma che coglie la presenza di eterogeneità non osservata; seppur significativo si passa dall'1 al 5 per mille forse ad indicare che parte di questa eterogeneità risiede proprio nel passato del processo di disoccupazione.

Per gli ultimi due modelli presentati sono state stimate le funzioni di rischio di base corrispondenti alla data specificazione del modello, fig. 4.6 .

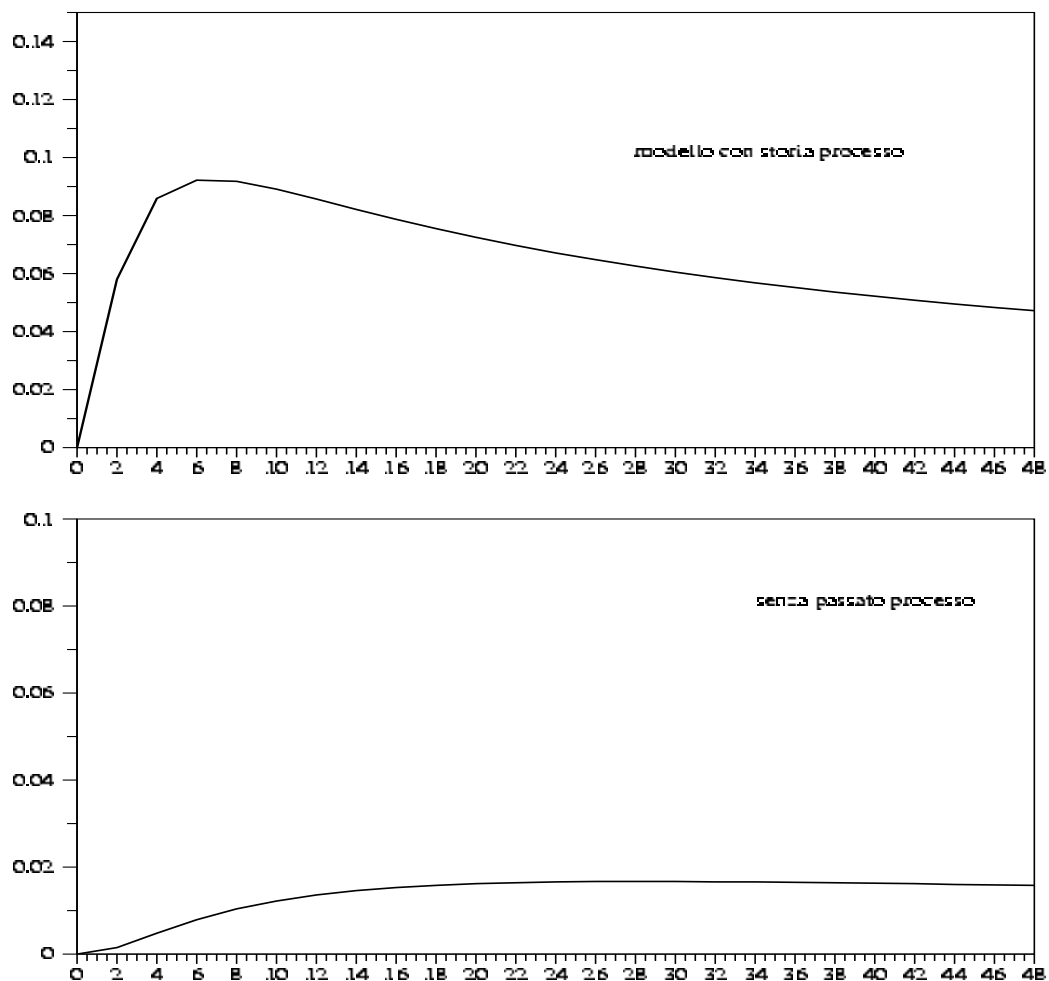


Figura 4.6: Funzioni di rischio risultanti dalle specificazioni nei modelli E.5- contenente la storia del processo- e E.4- senza le variabili riassuntive della storia del processo

Le funzioni stimate mostrano un andamento molto diverso tra loro indicando una diversa dipendenza temporale della transizione dalla durata osservata. La stima che si ottiene per il modello che contiene le variabili riassuntive di aspetti della storia passata del processo di disoccupazione mostra

una somiglianza maggiore con quella che si ottiene dalla stima della funzione con un metodo non parametrico nonchè alla forma che si ottiene specificando una funzione log-normale per il rischio di base, fig.4.2.

La stima di queste funzioni di base sembra mostrare che l'importanza della storia del processo risiede proprio nella determinazione della dipendenza temporale del tasso dalla durata: la funzione di transizione all'occupazione ha un'evoluzione temporale diversa se consideriamo lo stesso individuo con o senza la funzione riassuntiva dell'evoluzione della disoccupazione nel passato.



# Capitolo 5

## Conclusioni

Negli ultimi 20 anni si è assistito in Italia, e nel resto d'Europa, ad un considerevole aumento del tasso di disoccupazione generale e giovanile. Rispetto alla situazione europea l'Italia si caratterizza per un forte aumento non tanto delle persone disoccupate a causa della perdita di lavoro tanto per quelle che invece non riescono proprio ad entrare nel mercato del lavoro e che passano molto tempo nella condizione di "ricerca della prima occupazione". Tra questi troviamo, necessariamente, una forte percentuale di giovani che, usciti dal sistema formativo, si immettono o provano ad immettersi nel mercato del lavoro. Questo aumento si è verificato nonostante i livelli di qualifica dei giovani siano, nello stesso periodo, aumentati.

Questa situazione ha reso molto forte il dibattito relativo al fenomeno della disoccupazione, a quelli che sono i suoi riflessi sulla società nel suo complesso e a quelle che sono le possibili manovre da parte dello Stato in grado di alleviare la tensione sociale provocata da una quota sempre maggiore di disoccupati.

In una società nella quale il sistema economico si basa su redditi provenienti per la stragrande maggioranza da lavoro dipendente, l'essere disoccupato crea non poche problematiche a livello sociale e nei casi più estremi portano all'esclusione e alla povertà soprattutto in un sistema di welfare in cui la tutela del cittadino è direttamente legata alla sua posizione all'interno del Mercato del Lavoro.

Secondo le tipologie di welfare <sup>1</sup> evidenziate da Esping-Andersen [Esping-Andersen,1990] infatti, il sistema di welfare applicato in Italia rientra in grande misura tra

---

<sup>1</sup>Con welfare intendiamo l'insieme delle modalità di risposta ai bisogni emergenti della società; le tipologie sono state definite analizzando le interrelazioni tra le tre istituzioni fondamentali della vita sociale: il mercato del lavoro, la famiglia e lo stato. I differenti sistemi di welfare societari che distinguono i contesti nazionali, non sono altro, che il risultato delle modalità differenziate con le quali queste istituzioni si combinano

quelli di tipo " occupazionale-corporativo. In questo caso il sistema di welfare rappresenta un complemento del sistema economico, fornendo livelli di protezione legati allo status occupazionale. La rete di sicurezza sociale e l'acquisizione di diritti sociali sono, dunque, strettamente legati alla collocazione degli individui nel mercato del lavoro. La tendenza è verso una conservazione ed una cristallizzazione delle differenze di status, soprattutto attraverso un sostegno a modelli familiari tradizionali e prevalenti; vige, inoltre, il principio della sussidiarietà, con il soggetto pubblico che interviene solamente in caso di incapacità da parte della famiglia o delle reti parentali di far fronte alle proprie esigenze.

I trasferimenti monetari, piuttosto che l'erogazione di diretta di servizi alle famiglie, da parte di organizzazioni pubbliche o private, rappresentano un elemento essenziale del modello di welfare italiano. Le prestazioni di garanzia del reddito sono, in Italia ma anche all'interno dei modelli di stato sociale corporativo-conservatori dell'Europa continentale, connesse allo status occupazionale, ne risulta un grado di frammentazione delle prestazioni particolarmente accentuato, con la presenza di schemi differenziati sulla base della condizione lavorativa (dipendente o autonomo, pubblico o privato). Il meccanismo di garanzia del reddito favorisce soprattutto la tutela dei lavoratori collocati nei settori centrali del mercato del lavoro (grande industria, impiego pubblico), accreditando scarse, se non inesistenti, protezioni ai lavoratori appartenenti ai settori periferici.

In Italia manca inoltre un vero e proprio sussidio di disoccupazione e non si fornisce alcuna prestazione per i giovani che si trovano alla ricerca della prima occupazione.

La protezione dai rischi economici è connessa, nel nostro paese, alla garanzia di poter contare su un posto di lavoro possibilmente a tempo pieno ed indeterminato. Per chi fuoriesce da questa categoria, come i giovani, pochi, se non inesistenti, sono i meccanismi messi a punto per fornire un'adeguata rete di protezione sociale o di inserimento al lavoro ([La Rosa, Kieselbach, 1999]). Questa premessa ci mostra come il fenomeno della disoccupazione giovanile sia molto complesso e tiri in ballo diverse componenti del sistema " società" da quella puramente economica a quella sociale a quella demografica in quanto determina spesso un ritardo nei tempi in cui si vivono determinati eventi quali l'unione stabile o la nascita di un figlio.

L'aspetto affrontato in questo lavoro è quello della durata della disoccupazione con l'intento di verificare se e in quale verso questa possa essere influenzata dalle caratteristiche individuali ma soprattutto dalle caratteristiche del processo stesso di disoccupazione, intendendo con queste l'evoluzione dell'individuo rispetto al fenomeno nel corso del tempo.

Si è quindi cercato di identificare quali sono gli effetti di episodi precedenti

a quello attuale sulla probabilità che quest'ultimo si trasformi in un episodio lavorativo.

Come è stato introdotto nel capitolo 1 le carriere dei giovani nel mondo del lavoro sono caratterizzate da una forte instabilità iniziale che può durare anche per molti anni prima di essere assunti con un contratto che garantisca stabilità. Questa situazione si risolve nella presenza di molti episodi di occupazione, normalmente di breve durata, intervallati da episodi di disoccupazione di durata più o meno elevata.

Questa situazione è stata empiricamente riscontrata anche sul Data Base italiano dell'European Community Household Panel in cui per molti giovani tra i 16 e i 34 anni sono stati rilevati anche 10 episodi di disoccupazione nel corso di 5 anni. È qui evidente probabilmente l'effetto di azioni che risultano favorire l'occupazione a tempo determinato ma non si preoccupano di incentivarne poi l'assunzione a tempo indeterminato.

Sulla base dei dati a disposizione sono stati ricostruiti i percorsi individuali rispetto al mercato del lavoro, ovvero tutti gli ingressi e le uscite, e si è cercato di valutare quale fosse l'effetto di una " certa " traiettoria passata piuttosto che un'altra sulla probabilità che un episodio di disoccupazione osservato nel presente si trasformi in un lavoro.

La metodologia utilizzata è quella standard dei dati di durata e dei modelli di transizione che permettono la stima della probabilità condizionata di vivere una transizione da disoccupato a occupato e di legarla sia a delle covariate che alla durata osservata degli episodi.

Tali modelli sono stati affrontati utilizzando la teoria dei processi stocastici che rendono evidente come, nel momento in cui si stima un modello di transizione, si sta in realtà facendo un'ipotesi ben precisa sull'influenza che il passato del processo ha sulla sua evoluzione futura. La letteratura esistente sui modelli di transizione si basa normalmente su ipotesi di tipo markoviano o semi-markoviano. Nel primo caso l'influenza del passato è relegata allo stato di provenienza mentre nel secondo allo stato di provenienza ma anche alla durata trascorsa in tale stato.

L'obiettivo di questo lavoro è quello di verificare se il processo di disoccupazione giovanile non sia in realtà dominato, nelle sue dinamiche, da un processo stocastico più generale di quello semi-markoviano ovvero un processo nel quale l'influenza del passato si risolva in maniera più complessa che nei due tipi di processi citati.

Il lavoro è stato quindi rivolto alla ricerca di una funzione del passato del processo di disoccupazione che permettesse di rilasciare le ipotesi markoviane e semi-markoviane e di lavorare con processi più generali nei quali cioè si considera un'estensione maggiore del passato. La funzione ricercata considera alcuni aspetti del processo di disoccupazione-occupazione vissuto nel passato

e va a ritroso, fino a considerare alcune informazioni riguardanti i 5 anni precedenti la prima intervista oltre alla dinamica osservata attraverso la ricostruzione delle traiettorie individuali nel mercato del lavoro nel corso dei 5 anni di interviste del panel.

Tale funzione è stata utilizzata quale esplicativa dell'andamento e del livello del tasso di transizione assieme ad altre variabili rappresentanti caratteristiche individuali, familiari e di contesto al fine di valutare anche se il processo di disoccupazione giovanile fosse effettivamente un processo generale o non rispondesse invece a leggi di tipo markoviano o semi-markoviano: se così fosse la funzione del passato utilizzata dovrebbe risultare ininfluenza nella determinazione del rischio.

Va notato in ogni caso che ciò si potrebbe verificare anche per effetto della mancanza di informazioni ovvero per la scelta di aspetti non influenti; non sapendo valutare a priori quali potessero essere ci si è affidati alla reale disponibilità dei dati.

La funzione utilizzata prende in considerazione la durata dell'episodio precedente all'attuale sia di occupazione che di disoccupazione e le strategie utilizzate per la ricerca del lavoro stesso nel senso di ricerca più o meno attiva nonché la presenza nei 5 anni precedenti la prima intervista di episodi di disoccupazione, il loro numero e la presenza tra questi di episodi più lunghi di un anno .

Queste variabili sono risultate, nel complesso, importanti nel determinare i livelli del tasso di transizione degli individui indicando quindi che il passato del processo ha una sua importanza nel differenziare i livelli del tasso di transizione: a parità di altre condizioni l'aver sperimentato un percorso piuttosto che un altro nei 5 anni di osservazione risulta in una differenzialità dei livelli di rischio. In particolare osserviamo che le esperienze lavorative già vissute contribuiscono ad una diminuzione significativa della durata media dell'episodio attuale, avviene il contrario per la durata degli episodi di disoccupazione precedenti. Sembra quasi che le esperienze di precarietà che portano gli individui a sperimentare molti episodi di occupazione/disoccupazione possano essere positivi solo se costituiti da lunghi periodi di attività e brevi di inattività poiché, nel caso contrario, brevi lavori e lunghi periodi di attesa si ha in realtà una probabilità sempre minore di trovare lavoro. Tale funzione riassuntiva del passato del processo sembra comunque essere la determinante più importante della forma della dipendenza del rischio dalla durata: senza considerare l'evoluzione passata infatti la forma del rischio di base si discosta molto da quella stimata non parametricamente dall'insieme del campione ed è al contrario ad essa sovrapponibile nel momento in cui viene stimato il rischio di base prendendo esplicitamente in considerazione il passato.

La dipendenza temporale del rischio sembra quindi essere determinata in mo-

do rilevante dall'esperienza passata degli individui del campione. In conclusione possiamo affermare che la dipendenza temporale del rischio dalla durata non è governata da leggi semi-markoviane ma più generali ad indicare che l'esclusione del passato individuale dalla determinazione della probabilità di transizione al lavoro porterebbe ad interpretare in modo scorretto la forma del rischio di base e quindi la sua evoluzione nel tempo. La dipendenza temporale del rischio dalla durata risulta infatti in due forme differenti, per lo stesso individuo, quando si considera esplicitamente il suo passato o quando invece si ipotizza l'inesistenza di un passato differenziale. È solo nel primo caso però che la forma della funzione di rischio ricalca quella stimata non parametricamente sull'intero campione quasi a dimostrare che quest'ultima è in realtà fortemente determinata dall'esperienza degli individui presenti in esso.



# Appendice A

## Struttura del Personal file

### PERSONAL FILE

#### *P - GENERAL INFORMATION*

WAVE      WAVE NUMBER  
COUNTRY      COUNTRY CODE  
PID      PERSONAL IDENTIFICATION NUMBER  
HID      HOUSEHOLD IDENTIFICATION NUMBER  
PG001      BASIC HOUSEHOLD IDENTIFICATION NUMBER  
PG002      PERSONAL WEIGHT (CROSS-SECTIONAL)  
PG003      BASE WEIGHT  
PG004      NUMBER OF MINUTES TO COMPLETE THE INDIVIDUAL  
QUESTIONNAIRE  
PG005      DAY OF THE PERSONAL INTERVIEW  
PG006      MONTH OF THE PERSONAL INTERVIEW  
PG007      YEAR OF THE PERSONAL INTERVIEW

#### *P - DEMOGRAPHIC INFORMATION*

PD001      YEAR OF BIRTH OF THE INDIVIDUAL  
PD002      MONTH OF BIRTH OF THE INDIVIDUAL  
PD003      AGE OF THE INDIVIDUAL  
PD004      SEX OF THE INDIVIDUAL  
PD005      MARITAL STATUS OF THE PERSON  
PD006      DATE OF LAST CHANGE IN MARITAL STATUS  
PD007      PERSON IS LIVING IN CONSENSUAL UNION

#### *P - EMPLOYMENT*

PE001      MAIN ACTIVITY STATUS - SELF-DEFINED  
PE001A      MAIN ACTIVITY STATUS - SELF-DEFINED (A)  
PE002      MAIN ACTIVITY STATUS (REGROUPED) - SELF-DEFINED  
PE002A      MAIN ACTIVITY STATUS (REGROUPED) - SELF-DEFINED  
(A) PE003      ILO MAIN ACTIVITY DURING CURRENT YEAR

PE004 STATUS IN EMPLOYMENT  
 PE005 TOTAL HOURS WORKING /WEEK  
 PE005A HOURS WORKING / WEEK IN THE MAIN JOB  
 PE005B REASON FOR WORKING LESS THAN FULL TIME (LESS THAN 30 HOURS) IN MAIN JOB  
 PE005C MAIN JOB: FULL-TIME / PART TIME  
 PE006A CURRENT OCCUPATION - ANONYMISED (A)  
 PE006B CURRENT OCCUPATION - ANONYMISED (B)  
 PE006C CURRENT OCCUPATION - ANONYMISED (C)  
 PE007A CURRENT INDUSTRY - ANONYMISED (A)  
 PE007B CURRENT INDUSTRY - ANONYMISED (B)  
 PE007C CURRENT INDUSTRY - ANONYMISED (C)  
 PE008 SIZE OF LOCAL UNIT - CURRENT JOB  
 PE009 SECTOR OF THE BUSINESS OR ORGANIZATION - CURRENT JOB  
 PE010 JOB STATUS - CURRENT JOB (PERSONS IN PAID EMPLOYMENT AS MAIN ACTIVITY)  
 PE011 YEAR OF START OF CURRENT JOB  
 PE012 MONTH OF START OF CURRENT JOB  
 PE013 STEPS FOR FINDING THE CURRENT JOB  
 PE014 EXISTENCE OF AN UNEMPLOYMENT PERIOD BEFORE CURRENT JOB  
 PE015 NUMBER OF MONTHS OF UNEMPLOYMENT BEFORE CURRENT JOB  
 PE016 SKILLS OR QUALIFICATIONS TO DO A MORE DEMANDING JOB (FOR CURRENT JOB)  
 PE017 USE OF A FOREIGN LANGUAGE IN CURRENT JOB  
 PE018 FIRST FOREIGN LANGUAGE USED IN CURRENT JOB  
 PE019 SECOND FOREIGN LANGUAGE USED IN CURRENT JOB  
 PE020 THIRD FOREIGN LANGUAGE USED IN CURRENT JOB  
 PE021 SKILLS NEEDED FOR CURRENT JOB GIVEN BY A FORMAL TRAINING OR EDUCATION  
 PE022 IMPORTANCE OF THE TRAINING / EDUCATION TO PRESENT JOB  
 PE023 APPRENTICESHIP / TRAINING PAID BY EMPLOYER (WAGE) OR SOCIAL SECURITY (BENEFIT)  
 PE024 TYPE OF EMPLOYMENT CONTRACT  
 PE025 TOTAL LENGTH OF THE CONTRACT  
 PE026 CHILD-MINDING OR CRECHE PROVIDED BY EMPLOYER  
 PE027 HEALTH CARE OR MEDICAL INSURANCE PROVIDED BY EMPLOYER



PE028 EDUCATION OR TRAINING PROVIDED BY EMPLOYER  
 PE029 SPORTS AND LEISURE, HOLIDAY CENTER PROVIDED BY EMPLOYER  
 PE030 FREE OR SUBSIDISED HOUSING PROVIDED BY EMPLOYER  
 PE031 SATISFACTION WITH EARNINGS IN PRESENT JOB  
 PE032 SATISFACTION WITH JOB SECURITY IN PRESENT JOB  
 PE033 SATISFACTION WITH TYPE OF WORK IN PRESENT JOB  
 PE034 SATISFACTION WITH NUMBER OF WORKING HOURS IN PRESENT JOB  
 PE035 SATISFACTION WITH WORKING TIME IN PRESENT JOB  
 PE036 SATISFACTION WITH WORKING CONDITIONS/ENVIRONMENT IN PRESENT JOB  
 PE037 SATISFACTION WITH DISTANCE TO JOB / COMMUTING IN PRESENT JOB  
 PE038 NUMBER OF DAYS OF ABSENCE FROM CURRENT JOB DURING THE PAST 4 WEEKS BECAUSE OF ILLNESS OR OTHER REASON  
 PE039 AGE AT WHICH THE PERSON STARTED HIS/HER WORKING LIFE

***P - UNEMPLOYMENT***

PU001 PERSON HAS EVER BEEN UNEMPLOYED BEFORE FIRST JOB OR BUSINESS  
 PU002 PERSON HAS BEEN UNEMPLOYED DURING THE FIVE YEARS BEFORE JOINING THE SURVEY  
 PU002A PERSON HAS BEEN UNEMPLOYED AFTER 1989  
 PU003 NUMBER OF TIMES THE PERSON HAS BEEN UNEMPLOYED DURING THE FIVE YEARS BEFORE JOINING THE SURVEY  
 PU003A NUMBER OF TIMES THE PERSON HAS BEEN UNEMPLOYED AFTER 1989  
 PU004 IS ONE OF THE UNEMPLOYMENT SPELLS (DURING THE FIVE YEARS BEFORE JOINING THE SURVEY) LONGER THAN ONE YEAR  
 PU004 IS ONE OF THE UNEMPLOYMENT SPELLS (SINCE 1989) LONGER THAN ONE YEAR

***P - SEARCH FOR WORK***

PS001 LOOKING FOR JOB  
 PS001A MAIN REASON THE PERSON IS NOT SEEKING WORK  
 PS002 OPTIMUM NUMBER OF HOURS THE PERSON WOULD LIKE TO WORK IN THE NEW / ADDITIONAL WORK  
 PS003 CHANCES OF FINDING THE NEW / ADDITIONAL JOB  
 PS004 CONTACT WITH A PUBLIC EMPLOYMENT OFFICE ABOUT

FINDING WORK

- PS005 ACTIVE SEARCH TO FIND A JOB
- PS006 MAIN REASON NO ACTIVE STEPS TAKEN
- PS007 MINIMUM NET MONTHLY INCOME THE PERSON WOULD ACCEPT TO WORK
- PS008 AVAILABLE TO START WITHIN 2 WEEKS
- PS009 UNEMPLOYMENT BENEFIT OR ASSISTANCE RECEIVED AT PRESENT
- PS010 PERSON REGISTERED AT A PUBLIC EMPLOYMENT OFFICE FOR WORK
- PS011 PERSON RECEIVED A JOB OFFER DURING PAST 4 WEEKS
- PS012 ANY JOB OFFER ACCEPTED PS013 DATE OF START OF THE NEW JOB
- PS014 REASON FOR REFUSING THE MOST RECENT JOB OFFER

***P - PREVIOUS JOB***

- PJ001 PERSON HAS WORKED BEFORE
- PJ002 YEAR STOP WORKING IN LAST JOB
- PJ003 MONTH STOP WORKING IN LAST JOB
- PJ004 REASON FOR STOPPING IN PREVIOUS JOB
- PJ005 COMPARISON OF PRESENT AND PREVIOUS JOBS
- PJ006 STATUS IN EMPLOYMENT (PREVIOUS JOB)
- PJ007A OCCUPATION IN PREVIOUS JOB (ANONYMISED)
- PJ007B OCCUPATION IN PREVIOUS JOB (ANONYMISED)
- PJ007C OCCUPATION IN PREVIOUS JOB (ANONYMISED)
- PJ008A PREVIOUS INDUSTRY (ANONYMISED)
- PJ008B PREVIOUS INDUSTRY (ANONYMISED)
- PJ008C PREVIOUS INDUSTRY (ANONYMISED)
- PJ009 NUMBER OF EMPLOYEES IN THE LOCAL UNIT IN PREVIOUS JOB
- PJ010 PREVIOUS JOB IN PUBLIC / PRIVATE SECTOR
- PJ011 JOB STATUS - PREVIOUS JOB
- PJ012 FULL / PART TIME - PREVIOUS JOB

***P - CALENDAR OF ACTIVITIES LAST YEAR***

- PC001 MAIN ACTIVITY STATUS IN JANUARY
- PC002 MAIN ACTIVITY STATUS IN FEBRUARY
- PC003 MAIN ACTIVITY STATUS IN MARCH
- PC004 MAIN ACTIVITY STATUS IN APRIL
- PC005 MAIN ACTIVITY STATUS IN MAY
- PC006 MAIN ACTIVITY STATUS IN JUNE
- PC007 MAIN ACTIVITY STATUS IN JULY

PC008 MAIN ACTIVITY STATUS IN AUGUST  
PC009 MAIN ACTIVITY STATUS IN SEPTEMBER  
PC010 MAIN ACTIVITY STATUS IN OCTOBER  
PC011 MAIN ACTIVITY STATUS IN NOVEMBER  
PC012 MAIN ACTIVITY STATUS IN DECEMBER  
PC013 MOST FREQUENT ACTIVITY

***P - INCOME***

PI001 MAIN SOURCE OF PERSONAL INCOME  
PI010 WITHIN-HOUSEHOLD NON-RESPONSE INFLATION FAC-  
TOR  
PI100 TOTAL NET PERSONAL INCOME (DETAILED, NC, TOTAL  
YEAR PRIOR TO THE SURVEY)  
PI110 TOTAL NET INCOME FROM WORK (NET, NC, TOTAL YEAR  
PRIOR TO THE SURVEY)  
PI111 WAGE AND SALARY EARNINGS (NET, NC, TOTAL YEAR  
PRIOR TO THE SURVEY)  
PI1111 WAGE AND SALARY EARNINGS (REGULAR)  
PI1112 WAGE AND SALARY EARNINGS (LUMP SUM)  
PI112 SELF-EMPLOYMENT INCOME (NET)  
PI120 NON-WORK PRIVATE INCOME (NET, NC, TOTAL YEAR  
PRIOR TO THE SURVEY)  
PI121 CAPITAL INCOME  
PI122A ASSIGNED PROPERTY/RENTAL INCOME  
PI123 PRIVATE TRANSFERS RECEIVED  
PI130 TOTAL SOCIAL/SOCIAL INSURANCE RECEIPTS (NET, NC,  
YEAR PRIOR TO THE SURVEY)  
PI131 UNEMPLOYMENT RELATED BENEFITS  
PI132 OLD-AGE / SURVIVORS' BENEFITS  
PI1321 OLD-AGE RELATED BENEFITS  
PI132 2 SURVIVORS' BENEFITS  
PI133 FAMILY-RELATED ALLOWANCES  
PI134 SICKNESS/INVALIDITY BENEFITS  
PI135 EDUCATION-RELATED ALLOWANCES  
PI136 ANY OTHER (PERSONAL) BENEFITS  
PI137A ASSIGNED SOCIAL ASSISTANCE  
PI138A *qqquad*ASSIGNED HOUSING ALLOWANCE  
PI211M CURRENT WAGE AND SALARY EARNINGS (NET, MONTH-  
LY)

***P - TRAINING AND EDUCATION***

PT001 HAS THE PERSON BEEN IN EDUCATION OR TRAINING  
SINCE JANUARY 1995

PT002 WHICH KIND OF COURSE (S) WAS IT?  
 PT003 STARTING YEAR OF THE GENERAL EDUCATION COURSE  
 PT004 STARTING MONTH OF THE GENERAL EDUCATION COURSE  
 PT005 FINISHING YEAR OF THE GENERAL EDUCATION COURSE  
 PT006 FINISHING MONTH OF THE GENERAL EDUCATION COURSE  
 PT007 LEVEL OF THE GENERAL EDUCATION COURSE  
 PT007A LEVEL OF THE GENERAL EDUCATION COURSE (AG-  
 GREGATED)  
 PT008 STARTING YEAR OF THE VOCATIONAL EDUCATION COURSE  
 PT009 STARTING MONTH OF THE VOCATIONAL EDUCATION  
 COURSE  
 PT010 FINISHING YEAR OF THE VOCATIONAL EDUCATION COURSE  
 PT011 FINISHING MONTH OF THE VOCATIONAL EDUCATION  
 COURSE  
 PT012 LEVEL OF THE VOCATIONAL EDUCATION COURSE  
 PT013 OVERALL DURATION OF THE VOCATIONAL EDUCATION  
 COURSE  
 PT014 NUMBER OF DAYS OF THE VOCATIONAL EDUCATION  
 COURSE  
 PT015 NUMBER OF WEEKS OF THE VOCATIONAL EDUCATION  
 COURSE  
 PT016 NUMBER OF MONTHS OF THE VOCATIONAL EDUCA-  
 TION COURSE  
 PT017 WAS THE VOCATIONAL EDUCATION COURSE PAID FOR  
 OR ORGANIZED BY THE EMPLOYER  
 PT018 WAS THE VOCATIONAL EDUCATION COURSE A FULL-  
 TIME, A PART-TIME ATTENDANCE COURSE OR A COURSE BY COR-  
 RESPONDENCE  
 PT019 NUMBER OF HOURS PER WEEK IF THE VOCATIONAL  
 EDUCATION COURSE WAS A PART-TIME ATTENDANCE COURSE  
 PT020 PERSON TOOK THE VOCATIONAL EDUCATION COURSE  
 TO IMPROVE SKILLS OR JOB PROSPECTS  
 PT021 WAS IT USEFUL FOR THIS PURPOSE?  
 PT022 HIGHEST LEVEL OF GENERAL OR HIGHER EDUCATION  
 COMPLETED

***P - HEALTH***

PH001 HEALTH OF THE PERSON IN GENERAL  
 PH002 PERSON HAS ANY CHRONIC PHYSICAL OR MENTAL  
 HEALTH PROBLEM, ILLNESS OR DISABILITY  
 PH003 PERSON IS HAMPERED IN DAILY ACTIVITIES BY THIS  
 PHYSICAL OR MENTAL HEALTH PROBLEM, ILLNESS OR DISABILI-

TY (ONLY FOR PERSONS WITH THE PHYSICAL OR MENTAL HEALTH PROBLEM, ILLNESS OR DISABILITY)

PH003A PERSON IS HAMPERED IN DAILY ACTIVITIES BY ANY PHYSICAL OR MENTAL HEALTH PROBLEM, ILLNESS OR DISABILITY (ALL PERSONS)

PH004 HAS THE PERSON HAD TO CUT DOWN THINGS HE/SHE USUALLY DO ABOUT THE HOUSE, AT WORK OR IN FREE TIME BECAUSE OF ILLNESS OR INJURY?

PH005 HAVE YOU HAD TO CUT DOWN THINGS YOU USUALLY DO ABOUT THE HOUSE, AT WORK OR IN YOUR FREE TIME BECAUSE OF AN EMOTIONAL OR MENTAL HEALTH PROBLEM

PH006 PERSON ADMITTED IN AN HOSPITAL AS AN IN-PATIENT DURING THE PAST 12 MONTHS

PH007 NUMBER OF NIGHTS SPENT IN HOSPITAL DURING THE PAST 12 MONTHS

PH008 NUMBER OF TIMES THE PERSON HAS CONSULTED A GENERAL PRACTITIONER

PH009 NUMBER OF TIMES THE PERSON HAS CONSULTED A MEDICAL SPECIALIST

PH010 NUMBER OF TIMES THE PERSON HAS CONSULTED A DENTIST

PH011 NUMBER OF TIMES THE PERSON HAS BEEN TO A DOCTOR OR A DENTIST OR OPTICIAN (AGGREGATED)

PH012 PERSON ENTITLED TO FREE OR NEARLY FREE MEDICAL INSURANCE (IN HIS / HER NAME OR THROUGH ANOTHER FAMILY MEMBER)

PH013 PERSON IS COVERED BY PRIVATE MEDICAL INSURANCE

PH014 HOW IS THE PRIVATE MEDICAL INSURANCE PAID FOR

PH015 MONTHLY AMOUNT OF PRIVATE MEDICAL INSURANCE

***P - SOCIAL RELATIONS***

PR001 OTHER PERSONS PRESENT AT THE INTERVIEW

PR002 MEMBER OF A CLUB OR ORGANIZATION (SPORT OR ENTERTAINMENT CLUB, A LOCAL OR NEIGHBORHOOD GROUP, A PARTY)

PR003 FREQUENCY OF TALK TO THE NEIGHBORS

PR004 FREQUENCY OF MEETING PEOPLE (FRIENDS OR RELATIVES NOT LIVING WITH THE PERSON) AT HOME OR ELSEWHERE

PR005 SPOKEN TO SOMEONE OUTSIDE YOUR HOUSEHOLD (EVEN BY PHONE) LAST WEEK PR006 DAILY ACTIVITIES INCLUDE, WITHOUT PAY, LOOKING AFTER CHILDREN OR OTHER PERSONS

PR007 NUMBER OF HOURS (PER WEEK) SPENT LOOKING AF-

TER CHILDREN (GROUPED)  
PR007A NUMBER OF HOURS (PER WEEK) SPENT LOOKING AF-  
TER CHILDREN (DETAILED)  
PR008 NUMBER OF HOURS (PER WEEK) SPENT LOOKING AF-  
TER A PERSON OTHER THAN A CHILD (GROUPED)  
PR008A NUMBER OF HOURS (PER WEEK) SPENT LOOKING AF-  
TER A PERSON OTHER THAN A CHILD (DETAILED)  
PR009 LOOKED AFTER PERSON(S) (OTHER THAN CHILDREN)  
LIVE IN THE HOUSEHOLD OR ELSEWHERE  
PR010 LOOKING AFTER CHILDREN OR PERSON(S) (OTHER THAN  
CHILDREN) PREVENTS THE PERSON FROM UNDERTAKING THE  
AMOUNT OR KIND OF PAID WORK HE/SHE WOULD OTHERWISE  
DO

***P - MIGRATION***

PM001 MIGRATION TRAJECTORY  
PM002 NUTS CODE OF THE REGION WHERE THE PERSON IS  
CURRENTLY LIVING  
PM003 YEAR OF ARRIVAL IN THIS REGION  
PM005A CODE OF LAST FOREIGN COUNTRY OF RESIDENCE  
BEFORE COMING TO PRESENT COUNTRY (DETAILED - 11 MODAL-  
ITIES)  
PM005B CODE OF LAST FOREIGN COUNTRY OF RESIDENCE  
GROUPED BEFORE COMING TO PRESENT COUNTRY (GROUPED  
- 7 MODALITIES)  
PM005C CODE OF LAST FOREIGN COUNTRY OF RESIDENCE  
BEFORE COMING TO PRESENT COUNTRY (GROUPED - 2 MODAL-  
ITIES)  
PM006 YEAR OF ARRIVAL IN THE COUNTRY OF PRESENT RES-  
IDENCE  
PM007A CODE OF FOREIGN COUNTRY OF BIRTH (DETAILED -  
11 MODALITIES)  
PM007B CODE OF FOREIGN COUNTRY OF BIRTH (GROUPED -  
7 MODALITIES)  
PM007C CODE OF FOREIGN COUNTRY OF BIRTH (GROUPED -  
2 MODALITIES)  
PM008 CITIZENSHIP (ABBREVIATED)  
PM010 EXISTENCE OF A SECOND CITIZENSHIP  
PM011 SECOND CITIZENSHIP (ABBREVIATED)

***P - SATISFACTION***

PK001 SATISFACTION WITH WORK OR MAIN ACTIVITY  
PK002 SATISFACTION WITH FINANCIAL SITUATION

PK003 SATISFACTION WITH HOUSING SITUATION  
PK004 SATISFACTION WITH AMOUNT OF LEISURE TIME





# Appendice B

## Codifiche per la sezione Calendar of activities

Codes	Labels
1	Paid employment, whether full-time or part-time
2	Paid apprenticeship or training under special schemes related to employment
3	Self-employment
4	Unpaid work in family enterprise
5	In education or training
6	Unemployed
7	Retired
8	Doing housework, looking after children or other persons
9	In community or military service
10	Other economically inactive
-8	not applicable
-9	missing

Base: all persons with a completed personal questionnaire in wave i.

### Aggregazione dei codici del Calendario delle attività per la definizione degli stati

Stati	Codici corrispondenti
0	1, 2, 3, 4
1	6
2	5, 7, 8, 9



# Appendice C

## Matrice di transizione

Prime 20 linee della matrice di transizione originaria

PID	begin.month	end.month	old.state	new.state	length
5103	0	30	3	1	30
5103	30	33	1	3	3
5103	33	37	3	1	4
5103	37	48	1	1	11
6101	0	36	1	1	36
6102	0	36	3	-1	36
8103	0	12	1	-1	12
8104	0	12	2	-1	12
11103	0	12	1	-1	12
11103	37	48	1	1	11
12103	0	12	2	-1	12
14102	0	48	3	3	48
15103	0	24	3	1	24
15103	24	28	1	3	4
15103	29	36	3	2	8
15103	37	45	2	1	8
15103	46	60	1	1	2
15104	13	24	1	2	12
15104	25	48	2	2	23

Le linee rendono visibili gli episodi generati dallo stesso individuo.



## Appendice D

### Curve di sopravvivenza condizionate.

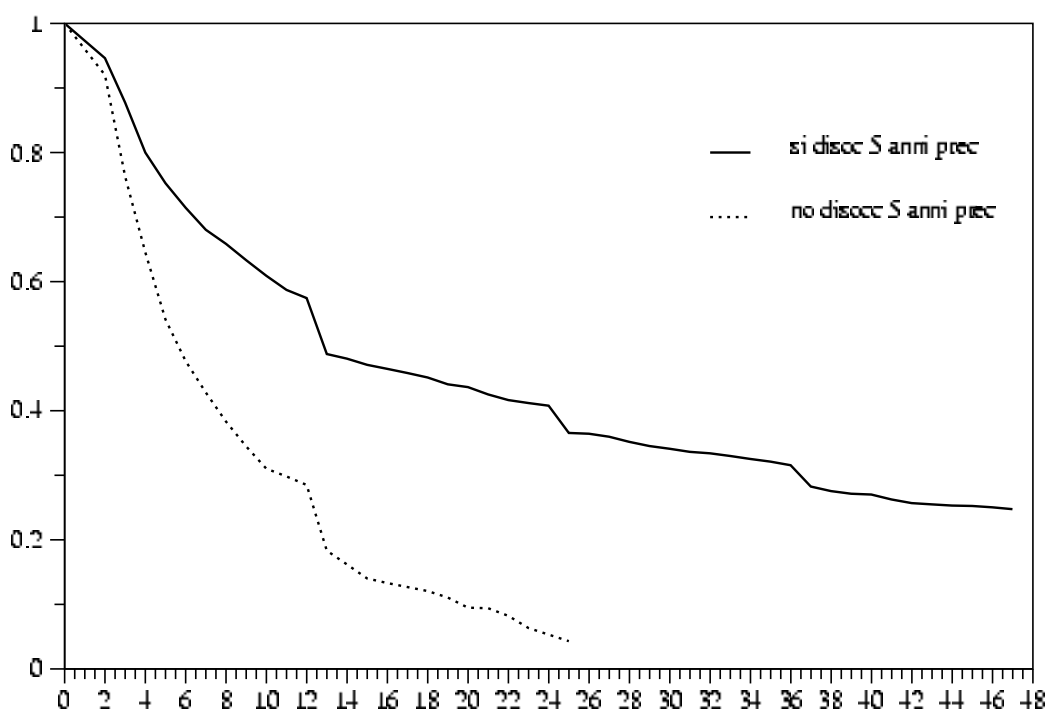


Figura D.1: Funzione di sopravvivenza condizionata all'essere o meno alla prima disoccupazione degli ultimi 5 anni precedenti il momento dell'intervista

test statistic	T-stat	DF	$\alpha$
Log-Rank (Savage)	134.6028	1	0.0000
Wilcoxon (Breslow)	107.1506	1	0.0000
Wilcoxon (Tarone-Ware)	120.4078	1	0.0000
Wilcoxon (Prentice)	115.3498	1	0.0000

Tabella D.1: Test statistici per il confronto delle funzioni di sopravvivenza della fig.D.1

test statistic	T-stat	DF	$\alpha$
Log-Rank (Savage)	89.2940	1	0.0000
Wilcoxon (Breslow)	86.7650	1	0.0000
Wilcoxon (Tarone-Ware)	91.9319	1	0.0000
Wilcoxon (Prentice)	88.9329	1	0.0000

Tabella D.2: Test statistici per il confronto delle funzioni di sopravvivenza di fig. D.2

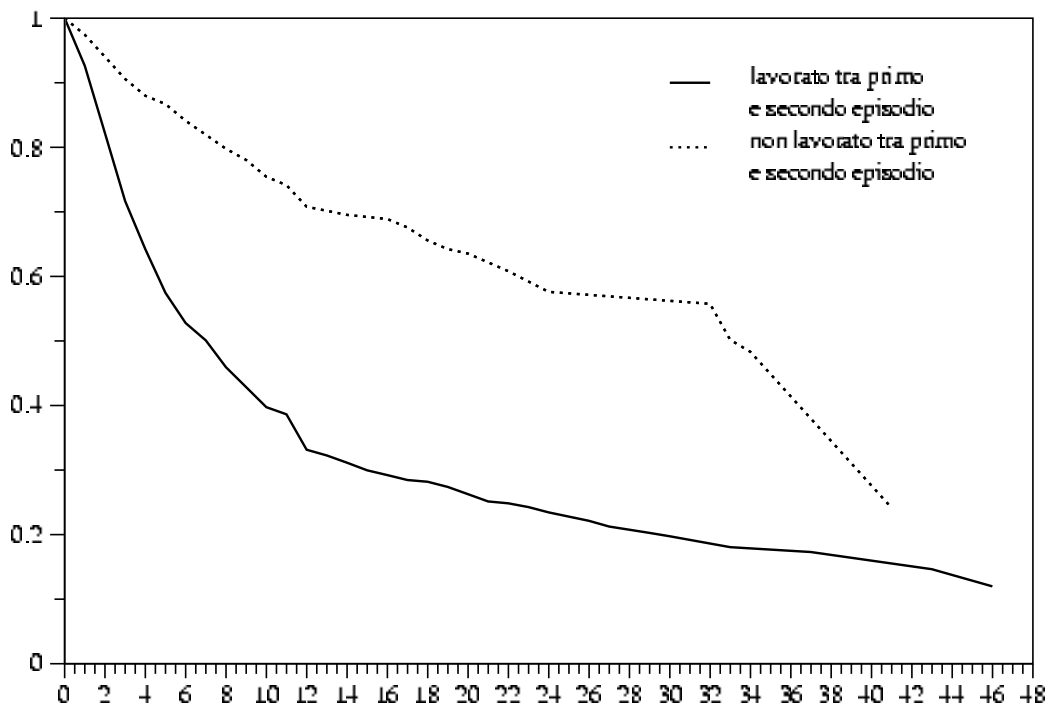


Figura D.2: Funzione di sopravvivenza condizionata all'aver lavorato tra primo e secondo episodio osservato di disoccupazione

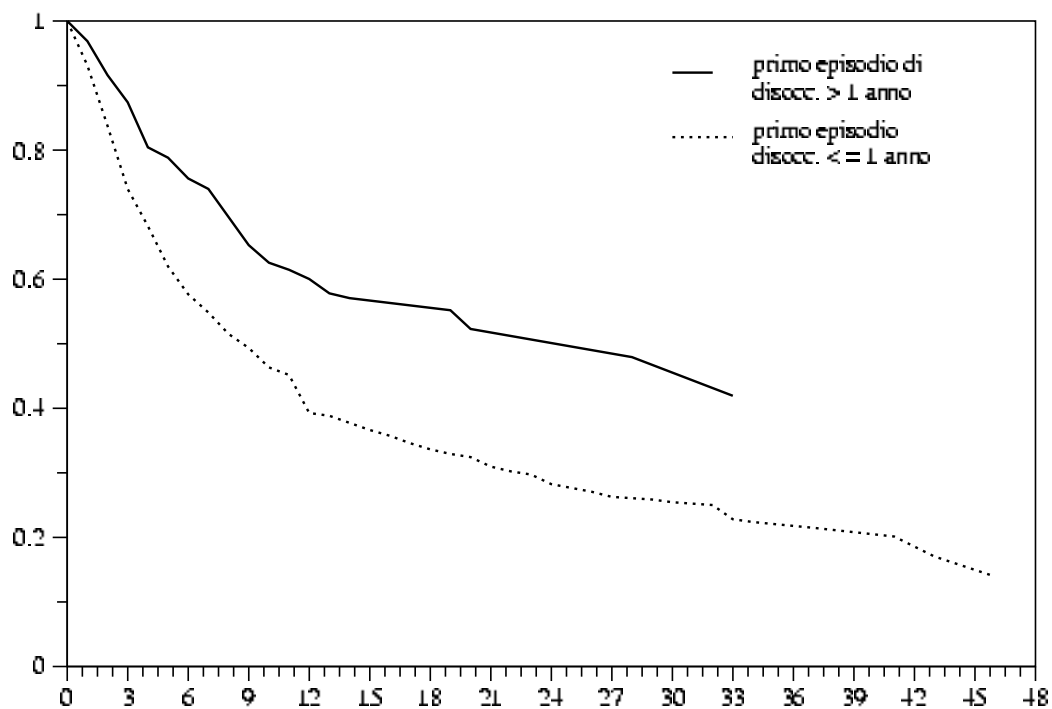


Figura D.3: Funzioni di sopravvivenza dei secondi episodi condizionati alla durata del primo

test statistic	T-stat	DF	$\alpha$
Log-Rank (Savage)	25.9840	1	0.0000
Wilcoxon (Breslow)	23.6332	1	0.0000
Wilcoxon (Tarone-Ware)	25.1884	1	0.0000
Wilcoxon (Prentice)	25.0385	1	0.0000

Tabella D.3: Test statistici per il confronto delle funzioni di sopravvivenza di fig. D.3



# Appendice E

## Risultati della stima dei modelli

variabile	(+/-) <i>exp(coeff.)</i>	std(coeff.)	$\alpha$
$\mu$	8.7337	0.0468	0.0000
SER2	(-)0.9594	0.0531	0.0000
SER3	(-)1.6285	0.0794	0.0000
SER4+	(-)1.8358	0.0848	0.0000
$\sigma$	0.6946	0.0374	0.0000
$\gamma$	3.4380	0.0795	0.0000
log-likelihood=-7562.1869			

Tabella E.1: MOD1. Stima dei parametri della funzione di rischio di base condizionata alla cardinalità dell'episodio. Il segno è quello dato da  $\hat{\beta}$

variabile	(+/-)exp(coeff.)	std(coeff.)	$\alpha$
$\mu$	11.91	0.0230	0.0000
SER2	(-)1.3772	0.0502	0.0000
SER3	(-)1.9197	0.0818	0.0000
SER4+	(-)2.1719	0.0901	0.0000
$\sigma$	0.9777	0.0171	0.0005
$\gamma$	2.7787	0.0473	0.0000
rif=NORD			
CENTRO	1.3267	0.0622	0.0000
SUD	2.2219	0.0508	0.0000
ISOLE	2.1541	0.0590	0.0000
rif= non paga l'affitto			
AFFITTO	1.0618	0.0225	0.0077
rif= COORTE1			
COORTE2	(-)1.0774	0.0643	0.2466
COORTE3	(-) 1.0665	0.0981	0.5113
FEMMINA	1.1016	0.0409	0.0002
rif= Non vive in coppia			
COPPIA	(-)1.3218	0.0547	0.0000
ANNI	(-)1.0227	0.0171	0.0001
rif= primo quartile			
REDIND2	(-)1.429	0.0545	0.0001
REDIND3	(-)1.5443	0.0861	0.0001
REDIND4	(-)1.7117	0.0643	0.0001
log-likelihood=-7252.3318			

Tabella E.2: MOD2. Stima i parametri del modello in tabella E.1 e quelli delle variabili che raccolgono caratteristiche individuali

variabile	(+/-)exp(coeff)	std(coeff.)	$\alpha$
$\mu$	16.78	0.2461	0.0000
SER2	(-)1.5752	0.0502	0.0000
SER3	(-)2.1093	0.0818	0.0000
SER4+	(-)2.7557	0.0902	0.0000
$\sigma$	1.0212	0.0171	0.0037
$\gamma$	2.7787	0.0534	0.0000
rif=NORD			
CENTRO	1.3042	0.0624	0.0000
SUD	2.0957	0.0534	0.0000
ISOLE	2.0248	0.0618	0.0000
rif= non paga l'affitto			
AFFITTO	1.0644	0.0226	0.0060
rif= COORTE1			
COORTE2	(-)1.0809	0.0643	0.3466
COORTE3	(-) 1.0632	0.0979	0.2113
FEMMINA	1.1168	0.0412	0.0010
rif= Non vive in coppia			
COPPIA	(-)1.2473	0.0640	0.0011
ANNI	(-)1.0196	0.0075	0.0011
rif= primo quartile			
REDIND2	(-)1.3924	0.0553	0.0021
REDIND3	(-)1.4641	0.0881	0.0021
REDIND4	(-)1.5485	0.1177	0.0001
rif= primo quartile			
REDFAM2	(-)1.0039	0.0568	0.0549
REDFAM3	(-)1.0267	0.0583	0.3499
REDFAM4	(-)1.1783	0.0665	0.0011
rif= no bimbi minori 12 anni in famiglia BIMBI	(-)1.1254	0.0591	0.0050
Ampiezza famiglia	(-)1.0665	0.0184	0.0005
log-likelihood=-7196.1892			

Tabella E.3: MOD3. Stima i parametri del modello in tabella E.2 e quelli delle variabili che raccolgono caratteristiche famigliari

variabile	(+/-)expcoeff.	std(coeff.)	$\alpha$
$\mu$	18.2056	7.6691	0.0267
SER2	(-)1.5840	0.0518	0.0000
SER3	(-)2.1180	0.0842	0.0000
SER4+	(-)2.8820	0.0934	0.0000
$\sigma$	1.0219	0.0174	0.0039
$\gamma$	2.5123	0.0897	0.0010
rif=NORD			
CENTRO	1.2893	0.0676	0.0000
SUD	2.1587	0.0512	0.0000
ISOLE	2.0077	0.0643	0.0000
rif= non paga l'affitto			
AFFITTO	1.1254	0.0245	0.0050
rif= COORTE1			
COORTE2	(-)1.1028	0.0687	0.2355
COORTE3	(-) 1.1192	0.1009	0.2431
FEMMINA	1.1452	0.0311	0.0001
rif= Non vive in coppia			
COPPIA	(-)1.2250	0.0678	0.0019
ANNI	(-)0.8976	0.6524	0.0124
rif= primo quartile			1
REDIND2	(-)1.0087	0.0556	0.0028
REDIND3	(-)0.9845	0.0885	0.0024
REDIND4	(-)1.5193	0.1177	0.0005
rif= primo quartile			1
REDFAM2	(-)1.0062	0.0570	0.0549
REDFAM3	(-)1.0336	0.0585	0.3510
REDFAM4	(-)1.1886	0.0668	0.0012
rif= no bimbi minori 12 anni in famiglia			
BIMBI	(-)1.1338	0.0595	0.0046
Ampiezza famiglia	(-)1.0709	0.0186	0.0002
Tasso disocc tot	(-)1.8455	0.0821	0.0001
PIL	1.1612	0.1023	0.0000
log-likelihood=-7026.1176			

Tabella E.4: MOD4. Stima i parametri del modello di tabella E.3 e quelli delle variabili di contesto economico

variabile	(+/-)expcoeff.	std(coeff.)	$\alpha$
$\mu$	2.0444	0.3788	0.0050
SER2	(-)1.6569	0.0481	0.0000
SER3	(-)2.3025	0.0792	0.0000
SER4+	(-)3.0219	0.0877	0.0000
$\sigma$	1.0575	0.0175	0.0024
$\gamma$	1.8912	0.0956	0.0100
rif=NORD			
CENTRO	1.2157	0.0585	0.0008
SUD	1.6865	0.0520	0.0000
ISOLE	1.6848	0.0588	0.0000
rif= non paga l'affitto			
AFFITTO	1.0370	0.0210	0.0822
rif= COORTE1			
COORTE2	1.0496	0.0687	0.0473
COORTE3	1.1685	0.580	0.0030
FEMMINA	1.0485	0.0383	0.2161
rif= Non vive in coppia			
COPPIA	(-)0.8623	0.0589	0.0194
ANNI	(-)0.1130	0.1256	0.0312
rif= primo quartile			1
REDIND2	(-)1.1306	0.0571	0.0030
REDIND3	(-)1.1037	0.0876	0.0251
REDIND4	(-)1.2218	0.1223	0.0041
rif= primo quartile			1
REDFAM2	(-)1.0242	0.0528	0.0349
REDFAM3	(-)1.0554	0.0442	0.0234
REDFAM4	(-)1.170	0.0619	0.0020
rif= no bimbi minori 12 anni in famiglia			
BIMBI	(-)1.1329	0.0548	0.0032
Ampiezza famiglia	1.0550	0.0172	0.0010
Tasso disocc tot	1.1091	0.0333	0.0001
PIL	0.9875	0.2089	0.0010
rif= nessun passo attivo			
ACTIVE1	1.2305	0.6767	0.2409
ACTIVE2	(-) 1.5201	0.1177	0.0001
ACTIVE3	(-) 1.4202	0.0700	0.0000
ACTIVE3	(-) 1.3959	0.0445	0.0000
DURPREOCC	(-)1.0186	0.0030	0.0000
DURPREDISOCC	1.0167	0.0041	0.0010
UNEM5P	1.179	0.0875	0.0051
NUMUNEM	1.2450	0.0232	0.0000
LONYEAR	1.0153	0.0436	0.1059
log-likelihood=-6709.2919			

Tabella E.5: MOD5. Stima i parametri del modello in tabella E.4 e



# Bibliografia

- [Billari,Ongaro,1999] Billari F., Ongaro F., (1999), *Lasciare la famiglia d'origine: quando e perchè*,in *Nuzialità e fecondità in trasformazione: percorsi e fattori del cambiamento* a cura di De Sandre P., Pinnelli A., Santini A., il Mulino, Bologna.
- [Blossfeld,Rohwer,1995] Blossfeld H-P., Rohwer G., (1995), *Technics of event history modeling. New approaches to causal analysis*, Laurence Erlbraun Associates Publishers, New Jersey.
- [Cahuzac et al., 1994] Cahuzac E., Mouchart M., Van Der Linden B., (1994) , *Examining the econom(etr)ic relevance of discretizing panel data: an application to the belgian labour market*, Labour, 8(3), 455-488.
- [Cordon,1997] Cordon, J., (1997), *Youth residential independence and autonomy: A comparative study*. Journal of Family Issues, 18, 576-607.
- [Cox,Snell,1968] Cox D. R., Snell E. J., (1968) , *A general definition of residuals*, Journal of The Royal Statistical Society, B(30), 248-275.
- [Esping-Andersen,1990] Esping-Andersen G., (1990), *The Tree Worlds of Welfare Capitalism*, Polity Press, Cambridge.
- [Eurostat,1992] Eurostat (1992), *Labour Force Survey. Results*, Bruxelles.
- [Eurostat,1996] Eurostat (1996), *Labour Force Survey. Results*, Bruxelles.
- [Eurostat,1996] Eurostat (1996), *the European Community household Panel (ECHP): Survey methodology and implementation.*, Volume 1, Office for Official Publications of the European Communities, Luxembourg.
- [Eurostat,1997] Eurostat (1997), *Youth in the European Union, From Education to Working Life*. The statistical office of the European Communities, Luxembourg.

- [Eurostat,1998] Eurostat (1998), *ECHP. Data dictionary and description variable*. The statistical office of the European Communities, Luxembourg.
- [Flinn,Heckmann,1982] Flinn C., Heckmann J.J. (1982), *Models for the analysis of labor force dynamics*, in Basmann R., Rhodes C., *Advances in econometrics* Vol.1, JAI Press, Greenwich, pp.35-95.
- [Flinn,Heckmann,1983] Flinn C., Heckmann J.J. (1983), *Are unemployment and Out of labor force behaviorally distinct labor force states?*, *Journal of labor economics*,1, 29-42.
- [Florens et al., 1994] Florens J.P., Fougere D., Kamioka T., Mouchart M., (1994), *La modlization conomtric des transitions individuelles sur le march du travail*. *Economie et Prvision*, 116, 181-217.
- [Florens et al.,1995] Florens J.P., Fougre D.,Mouchart M., (1995), *Duration models*, Core Discussion Paper.
- [Heckman,Singer,1984a] Heckman J.J., Singer B. (1984a), *Econometric duration analysis*. *Journal of Econometrics*, 24, 63-132.
- [Heckman,Singer,1984b] Heckman J.J., Singer B. (1984b), *Longitudinal analysis of labor market data*, *Econometric Society Monographs* n10, Cambridge University Press.
- [HwaJung,Donqgyun,2002] HwaJung C., Donqgyun S., (2002), *Do past unemployment spells affect the duration of current unemployment?*, *Economics Letters*, 77, 157-161.
- [Kalbfleish,Prentice,1980] Kalbfleish J., Prentice R. (1980), *The statistical analysis of failure time data*, Wiley, New York.
- [Istat,1997] Istat, (1997), *Rapporto sull'Italia*, il Mulino, Bologna.
- [Istat,1999] Istat, (1999) *Rapporto sull'Italia*, Roma.
- [Lancaster,1990] Lancaster T. (1990), *The econometric Analysis of Transition Data*,Cambridge University Press, Cambridge.
- [La Rosa,Kieselbach,1999] La Rosa M., Kieselbach T, a cura di (1999), *Disoccupazione giovanile ed esclusione sociale*,FrancoAngeli, Milano.
- [Lawless,1982] Lawless J.F. (1982), *Statistical models and methods for lifetime data*, Wiley, New York.



- [Mouchart,2003] Mouchart M. (2003), *The econometrics of duration data and of point processes*, Lectures note.
- [OECD,1998] OECD (1998), *Getting started, settling in; The transition from education to the labour market*, in Employment Outlook, Paris.
- [OECD,1999] OECD (1999), *Employment in Europe*
- [Peracchi,2000] Peracchi F. (2000), *The European Community Household Panel: A review*. Empirical Economics, 27, 63-90.
- [Peracchi,Nicoletti,2002] Peracchi F.,Nicoletti C. (2002), *A cross-country comparison of survey participation in the ECHP*. Working Paper.
- [Rohwer,Pätter,1995] Rohwer G., Pätter U. (1995), *TDA-Transition Data Analysis- User's Manual*, Ruhr-Universität,<http://steinhaus.stat.ruhr-uni-bochum.de>, Bochum, Germany.