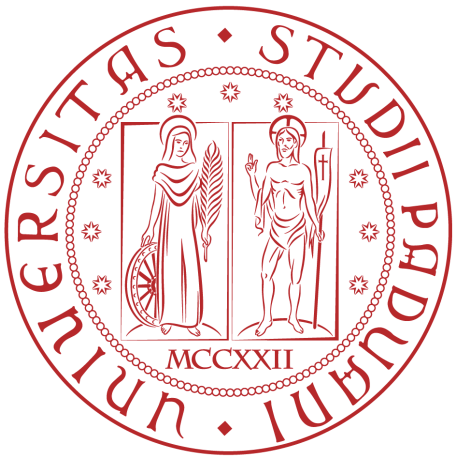
Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche



Una analisi del tempo intercorso tra il completamento dell'istruzione e la nascita del primo figlio in Europa.

Relatore Prof. Omar Paccagnella
Dipartimento di Scienze Statistiche

Laureanda Annachiara Dal Colle
Matricola N 1179896

Anno Accademico 2020/2021

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Per *fecondità* si intende generare figli nati vivi nei limiti biologici del periodo di fertilità delle donne. Il livello di fecondità di sostituzione, invece, sottintende il numero di figli per donna necessario a garantire il rimpiazzo delle generazioni dei genitori. Nei Paesi Occidentali a bassa mortalità, la fecondità di sostituzione ammonta a un numero appena superiore a due.

Come sostiene Zanier (2002), il progressivo declino della fecondità nei Paesi Occidentali costituisce un'evidenza nota e ampiamente dibattuta nella letteratura scientifica internazionale. Nella maggior parte dei Paesi dell'Europa Occidentale, l'età media alla maternità ha subìto un incremento di circa due anni (da 26 a 28) dalle generazioni del dopoguerra fino a quelle della prima metà degli anni Sessanta.

Se, da un lato, un aumento dell'età alla prima gravidanza è una caratteristica tipica delle recenti tendenze della fertilità nei Paesi sviluppati, dall'altro, non c'è ancora una chiara spiegazione di questo cambiamento sistematico e pervasivo.

Tra le cause potenziali, la crescita della partecipazione all'istruzione è spesso menzionata, ma le prove esplicite del suo ruolo come motore di questo cambiamento sono scarse (Ni Bhrolchain and Beaujouan, 2012).

A tal proposito, l'obiettivo della presente tesi è di indagare sul ruolo dell'istruzione nella fecondità. In particolare, la tesi sarà focalizzata sull'analisi del tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio in Europa.

Per l'obiettivo in questione, si analizzeranno i dati delle indagini SHARE e SHARELIFE. Il progetto SHARE ha prodotto un database che raccoglie dati di tipo longitudinale riguardanti la salute, lo stato socio-economico

e sociale e altre informazioni legate alle abitudini e agli stili di vita degli ultracinquantenni in Europa (Börsch-Supan et al., 2013). SHARELIFE, è un'indagine che raccoglie informazioni dettagliate sull'intera biografia degli intervistati SHARE. Entrambe le indagini sono realizzate per in moltissimi Paesi facenti parte dell'Unione Europea, più Svizzera e Israele.

L'indagine SHARELIFE, unita all'indagine SHARE, offre molteplici dati per provare a rispondere all'obiettivo del presente studio.

Nel primo capitolo viene introdotta, dal punto di vista teorico, la fecondità in Europa e i fattori che influenzano il cambiamento demografico di un Paese. Inoltre, viene presentato il progetto SHARE, ovvero l'Indagine sulla Salute, l'Invecchiamento e il Pensionamento in Europa, nonchè gli obiettivi del presente studio in maggior dettaglio.

Nel secondo capitolo vengono presentati i dati utilizzati e le relative analisi esplorative e descrittive.

Nel terzo capitolo verranno introdotti, dal punto di vista teorico, i dati di durata e i relativi modelli per studiare la sopravvivenza. Viene proposto, inoltre, il software R utilizzato per effettuare l'analisi statistica.

Nel quarto capitolo vengono presentate le applicazioni ai dati dei modelli per dati di durata. Infine, si conclude con una discussione sui risultati ottenuti.

L'ultimo capitolo illustra e riassume le principali conclusioni del presente lavoro.

Capitolo 1

La fecondità in Europa

1.1 Introduzione

Nel primo capitolo viene introdotta, dal punto di vista teorico, la fecondità in Europa e i principali fattori che influenzano il cambiamento demografico di un Paese. In particolare, verrà evidenziato il ruolo dell'istruzione nella scelta di avere dei figli.

Inoltre, viene presentato il progetto SHARE, ovvero l'Indagine sulla Salute, l'Invecchiamento e il Pensionamento in Europa (Börsch-Supan et al., 2013).

Infine, vengono presentati gli obiettivi di questo studio.

1.2 Fecondità e transizione demografica

La scelta di formare una coppia e avere dei figli è un aspetto chiave dei processi di riproduzione, in quanto permette a una popolazione di esistere nel tempo. La fecondità, in evoluzione nel corso degli anni, è fortemente legata ai cambiamenti sociali ed economici di un Paese.

Nella seconda metà del XX secolo, è avvenuta una riduzione della fecondità e si sono modificate le strutture e le relazioni familiari rispetto agli anni precedenti, che possono «derivare da una transizione verso società fondate su nuovi valori e nuove preferenze, così come da cambiamenti strutturali nei

livelli di istruzione e di partecipazione al mercato del lavoro» (Cantalini, 2017).

Si sono verificati profondi mutamenti di tipo demografico: in tutti i Paesi sviluppati sono diminuiti il numero dei matrimoni e sono aumentati divorzi e convivenze. Inoltre, è stata riscontrata una forte diminuzione del numero di nascite, tanto che si parla di "seconda transizione demografica".

La seconda transizione demografica (Van de Kaa, 1987) è una nuova rivoluzione demografica che segue, ma ha le sue radici, nella prima transizione demografica. Come nella prima, si è prodotto un disequilibrio tra i tassi di natalità e quelli di mortalità. A differenza della prima, questa nuova transizione è stata attivata dal declino della fecondità. In particolare, la fecondità è caduta da livelli in grado di garantire l'equilibrio demografico, a livelli ben al di sotto di 2,1 figli per donna, cioè al di sotto della fecondità di sostituzione[1], ed è questa la caratteristica principale di questo nuovo stadio dello sviluppo demografico.

Uno dei fattori che contribuisce ad abbassare il tasso di fecondità totale[2] di un Paese è l'infecondità, ovvero la quota di donne che restano senza figli al termine dalla vita riproduttiva. L'aumento dell'infecondità è diventato un fenomeno diffuso: il mutamento di alcuni valori sociali, economici e culturali, ha in qualche modo spinto le donne a non avere figli, o a ritardare la maternità (Sobotka, 2013).

I tassi di fecondità sono diminuiti costantemente dalla metà degli anni Sessanta fino alla fine del secolo scorso negli Stati membri dell'UE. Tuttavia, all'inizio degli anni 2000, il tasso di fecondità totale ha mostrato segni di nuovo aumento. Tale andamento si è arrestato nel 2010 e un successivo calo è stato osservato nel 2013, seguito da un lieve aumento verso il 2017 (European Commission, 2015).

[1] La fecondità di sostituzione, in un'ottica generazionale, è il tasso di fecondità che assicura ad una popolazione la possibilità di riprodursi mantenendo costante la propria struttura.

[2] Il tasso di fecondità totale (TFT) esprime il numero medio di figli per donna in età feconda (15-49 anni).

Total fertility rate, 1960–2018
(live births per woman)

	1960	1970	1980	1990	2000	2001	2010	2016	2017	2018
EU-27 (¹)	:	:	:	:	:	1.43	1.57	*1.57*	*1.56*	*1.55*
Belgium	2.54	2.25	1.68	1.62	1.67	1.67	1.86	1.68	1.65	1.62
Bulgaria	2.31	2.17	2.05	1.82	1.26	1.21	1.57	1.54	1.56	1.56
Czechia	2.09	1.92	2.08	1.90	1.15	1.15	1.51	1.63	1.69	1.71
Denmark	2.57	1.95	1.55	1.67	1.77	1.74	1.87	1.79	1.75	1.73
Germany	:	:	:	:	1.38	1.35	1.39	1.60	*1.57*	*1.57*
Estonia	1.98	2.17	2.02	2.05	1.36	1.32	1.72	1.60	1.59	1.67
Ireland	3.78	3.85	3.21	2.11	1.89	1.94	2.05	1.81	1.77	1.75
Greece	2.23	2.40	2.23	1.39	1.25	1.25	1.48	1.38	1.35	1.35
Spain	:	:	2.22	1.36	1.22	1.23	1.37	1.34	1.31	1.26
France	:	:	:	:	1.89	1.90	2.03	*1.93*	*1.90*	*1.88*
Croatia	:	:	:	:	:	1.46	1.55	1.42	1.42	1.47
Italy	2.37	2.38	1.64	1.33	1.26	1.25	1.46	1.34	1.32	1.29
Cyprus	:	:	:	2.41	1.64	1.57	1.44	1.37	1.32	1.32
Latvia	:	:	:	:	1.25	1.22	1.36	1.74	1.69	1.60
Lithuania	:	2.40	1.99	2.03	1.39	1.29	1.50	1.69	1.63	1.63
Luxembourg	2.29	1.97	1.50	1.60	1.76	1.66	1.63	1.41	1.39	1.38
Hungary	2.02	1.98	1.91	1.87	1.32	1.31	1.25	1.53	1.54	1.55
Malta	:	:	1.99	2.04	1.68	1.48	1.36	1.37	1.26	1.23
Netherlands	3.12	2.57	1.60	1.62	1.72	1.71	1.79	1.66	1.62	1.59
Austria	2.69	2.29	1.65	1.46	1.36	1.33	1.44	1.53	1.52	1.47
Poland (²)	:	:	:	2.06	1.37	1.31	1.41	1.39	1.48	1.46
Portugal	3.16	3.01	2.25	1.56	1.55	1.45	1.39	1.36	1.38	1.42
Romania	:	:	2.43	1.83	1.31	1.27	1.59	1.69	1.71	1.76
Slovenia	:	:	:	1.46	1.26	1.21	1.57	1.58	1.62	1.60
Slovakia	3.04	2.41	2.32	2.09	1.30	1.20	1.43	1.48	1.52	1.54
Finland	2.72	1.83	1.63	1.78	1.73	1.73	1.87	1.57	1.49	1.41
Sweden	:	1.92	1.68	2.13	1.54	1.57	1.98	1.85	1.78	1.76
United Kingdom	:	:	1.90	1.83	1.64	1.63	1.92	1.79	1.74	1.68
Iceland	:	2.81	2.48	2.30	2.08	1.95	2.20	1.74	1.71	1.71
Liechtenstein	:	:	:	:	1.57	1.52	1.40	1.61	1.44	1.58
Norway	:	2.50	1.72	1.93	1.85	1.78	1.95	1.71	1.62	1.56
Switzerland	2.44	2.10	1.55	1.58	1.50	1.38	1.52	1.54	1.52	1.52
Montenegro	:	:	:	:	:	:	1.70	1.79	1.78	1.76
North Macedonia	:	:	:	:	1.88	1.73	1.56	1.50	1.43	1.42
Albania	:	:	:	:	:	:	1.63	1.54	1.48	1.37
Serbia	:	:	:	:	1.48	1.58	1.40	1.46	1.49	1.49
Turkey	:	:	:	:	:	:	2.04	2.11	2.07	1.99

(¹) 2010, 2015 and 2017: break in series.
(²) 2000 and 2010: break in series.
Source: Eurostat (online data code: demo_frate)

eurostat

Figura 1.1: Tasso di fecondità totale in Europa, 1960-2018. Fonte: *Eurostat*

Nel 2018 il tasso di fecondità totale nell'UE-27 era di 1.55 nati vivi per donna e, come si può vedere in Figura 1.1, la Francia ha registrato il tasso di fecondità totale più elevato con 1.88 nati vivi per donna, seguita da Svezia, Romania e Irlanda. Per contro, i tassi di fecondità totali più bassi nel 2018 sono stati registrati a Malta (con 1.23 nati vivi per donna), Spagna (1.26 nati vivi per donna), Italia (1.29 nati vivi per donna).

I fattori demografici che possono influire sulla fecondità di un Paese sono molteplici.

In primo luogo, un ruolo importante viene svolto dal matrimonio. In particolare, le donne non sposate, sono quelle che hanno probabilità maggiore di non avere figli (Tanturri, 2009). Inoltre, considerando che il matrimonio è una forma d'unione molto diffusa, si può ipotizzare che anche l'aumento dell'età media alle nozze sia correlato positivamente con l'infecondità.

In secondo luogo, in Europa, l'età media alla nascita del primo figlio è di 29.3 anni (European Commission, 2015) e non sempre questo dato deriva

da un ritardo nella formazione della coppia. Ritardare la nascita del primo figlio può causare un abbassamento della fertilità, facendo insorgere problemi di sterilità individuali e/o di coppia. Si tratta della cosiddetta infecondità involontaria (Rowland, 2007): la gravidanza viene ritardata fino al punto in cui essa diviene improbabile o impossibile, nel qual caso il rinvio volontario viene trasformato in infecondità involontaria.

Un ulteriore fattore chiave è l'età all'uscita dalla casa dei genitori. Come evidenzia l'European Commission (2018b), i giovani tardano sempre di più ad uscire dalla famiglia d'origine, restandoci anche fino ai trentanni di età. Questo fenomeno può avere un'influenza sul tempo disponibile per creare una famiglia e di conseguenza generare infecondità non volontaria.

Inoltre, la crescita dell'indipendenza economica e dei livelli di istruzione tra le donne sono fattori che hanno avuto un forte peso nei cambiamenti avvenuti nella seconda transizione demografica. Questo aspetto verrà approfondito nel paragrafo 1.2.1.

Oltre ai precedenti fattori, va aggiunta la "rivoluzione di genere" in atto negli ultimi decenni, ovvero la caduta delle convenzioni sul ruolo delle donne – sulle quali storicamente ha gravato la totale quantità del lavoro domestico e di cura dei figli – e la spinta verso la parità di genere, non solo in campo socio-economico, ma anche in ambito familiare. Studi recenti (Arpino et al., 2015) hanno dimostrato che nei Paesi in cui la spinta verso l'equità di genere è stata più forte, si è avuta un'inversione nei trend di fecondità, che hanno ripreso a crescere, seppur in modo contenuto, nell'ultimo decennio. Allo stesso modo, un aumento della *gender equality* può essere correlata anche a una diminuzione dell'infecondità.

1.2.1 L'istruzione e la nascita di un figlio

Per le donne, la maternità è un'esperienza significativa e le sue tempistiche incidono nelle traiettorie di vita di una donna.

La ricerca di Miller (2011) evidenzia come l'età al primo figlio sia legata a diversi fattori, quali l'istruzione, il percorso professionale, i redditi e la salute delle donne.

In generale, l'istruzione, la condizione socio-economica e il reddito sono indicatori importanti del benessere in un adulto. Tra questi indicatori, l'istruzione è particolarmente importante perché porta a benefici per tutta la vita. Ad esempio, un maggiore livello di istruzione è collegato ad un maggior benessere in età adulta, in termini di guadagno economico e di salute (Pascarella and Terenzini, 2005).

Negli ultimi decenni i livelli di istruzione sono aumentati in modo significativo in molti Paesi, così come è aumentata l'età alla nascita del primo figlio. Un numero sempre crescente di donne ha raggiunto livelli di istruzione secondaria, superando quelle con istruzione primaria o inferiore, e anche l'istruzione universitaria è diventata sempre più comune, seppure con un andamento più lento (Beaujouan et al., 2015). La rapida crescita dei livelli di istruzione delle donne e l'aumento dell'età in cui hanno il loro primo figlio sono fattori chiave di transizione demografica per un paese (Gangadharan and Maitra, 2003).

Caldwell (1980) sostiene che la formazione scolastica sia una componente determinante di questo processo perché l'aumento del livello di istruzione influisce in modo significativo sia sull'età al matrimonio, sia sul tempo che intercorre tra il matrimonio e la nascita del primo figlio – in particolare aumentando sia l'età al matrimonio, sia il tempo per il primo figlio. La ricerca di Gangadharan and Maitra (2003) in Pakistan, conferma questa teoria e mostra che il livello di istruzione raggiunto da una donna ha un effetto significativo sull'età al matrimonio. Al contrario, però, la ricerca non evidenzia alcun effetto significativo del livello di istruzione - di entrambi i partner - sulla durata tra il matrimonio e la nascita del primo figlio.

Si sostiene che una maggiore istruzione favorisca «alternative economiche» allo sposarsi e avere figli. Un livello di istruzione più elevato potrebbe significare che, per le donne, il vantaggio dell'essere single potrebbe superare quello dell'essere sposate (Becker, 1974).

Inoltre, è probabile che le donne riducano la loro partecipazione al mercato del lavoro dopo la gravidanza e, di conseguenza, per una donna più istruita, il costo-opportunità di una gravidanza potrebbe essere più alto.

Allo stesso modo, anche l'utilità di ritardare la nascita potrebbe superare l'utilità di avere un figlio subito.

Come già menzionato nel paragrafo precedente, l'età al primo figlio è significativa nel processo di transizione demografica perché una maggiore età alla prima nascita è tipicamente associata a un tasso di fertilità più basso nel corso della vita.

Generalmente ci si aspetta un rapporto di correlazione positiva tra livello di istruzione e infecondità di coorte, anche a causa della forte evidenza di un'associazione positiva tra l'infecondità e l'istruzione a livello individuale (Nicoletti and Tanturri, 2008). Tuttavia, altri studi (Beaujouan et al., 2015) dimostrano che il crescente livello di istruzione potrebbe aver giocato un ruolo minore sull'infecondità perché i cambiamenti hanno influenzato la formazione della famiglia a tutti i livelli di istruzione. In alcune regioni europee, ad esempio nei Paesi Scandinavi, si riscontra un livello inferiore di infecondità nelle donne più istruite: Persson (2010) parla di "recupero di fecondità", ovvero la tendenza a diventare madri ad un'età più adulta nelle donne con titolo di studio più elevato.

1.3 L'indagine SHARE e SHARELIFE

SHARE (Survey of Health, Ageing and Retirement in Europe) è un acronimo per indicare l'Indagine sulla Salute, l'Invecchiamento e il Pensionamento in Europa. SHARE è un progetto di ricerca che aiuta a comprendere meglio le dinamiche di invecchiamento della popolazione.

Il progetto SHARE ha portato alla realizzazione di un database che raccoglie dati di tipo longitudinale riguardanti la salute, lo stato socio-economico e sociale e altre informazioni legate alle abitudini e agli stili di vita degli ultracinquantenni in Europa. (Börsch-Supan et al., 2013)

L'obiettivo principale di SHARE è quello fornire dati accurati per la ricerca sull'invecchiamento, attraverso la combinazione di

- transdisciplinarietà, studiando cioè le interazioni tra fattori bio-medici e socio-economici;

- longitudinalità, combinando dati longitudinali prospettici e retrospettivi;

- copertura europea e utilizzo di strumenti e metodologie di indagine armonizzati. Tutti i Paesi, infatti, hanno lo stesso programma di ricerca, somministrano un uguale questionario con le medesime modalità di intervista. Inoltre, la raccolta dei dati e i tassi di risposta di tutti i Paesi sono monitorati centralmente (Bergmann et al., 2019).

Gli intervistati dalla prima alla settima rilevazione, dette anche *waves*, sono complessivamente 140.000. Le interviste vengono svolte mediante la modalità CAPI (Computer Assisted Personal Interview) e l'autocompilazione di questionari cartacei alla fine dell'intervista CAPI.

L'indagine SHARE è svolta con cadenza biennale; ha avuto inizio nel 2004 con la prima wave e comprendeva dodici Paesi partecipanti, per arrivare alla settima wave del 2017, per un totale di 27 Paesi membri dell'Unione Europea, più Svizzera ed Israele.

Nella settima ondata di raccolta dati vengono combinati due diversi questionari: il questionario classico SHARE e un questionario che ripercorre l'intera vita delle persone, che prende il nome di SHARELIFE.

L'approccio classico di SHARE tiene traccia delle stesse persone, nel corso del tempo, dai 50 anni in su. Ad ogni rilevazione (ad eccezione della wave 3), gli intervistati rispondono al questionario SHARE, che raccoglie informazioni sulla loro vita in quel preciso momento. SHARE, quindi, documenta come gli intervistati reagiscono alle stesse domande e misurazioni nelle diverse ondate di rilevazione.

Tuttavia, rilevando le informazioni a partire dal cinquantesimo anno di età, le esperienze vissute negli anni precedenti non sono disponibili nei dati SHARE. Ciò rende difficile, per i ricercatori, la contestualizzazione di quanto rilevato.

Per sopperire a questa mancanza, nella wave 3 viene introdotto il questionario SHARELIFE, il quale si focalizza sull'intera biografia degli intervistati SHARE: ai rispondenti viene infatti chiesto di collocare nel tempo i prin-

cipali eventi della propria vita. Questa prospettiva è particolarmente utile per l'analisi di effetti a lungo termine, come, ad esempio, gli effetti delle condizioni di infanzia sul benessere in età adulta.

Nella settima wave, l'intervista SHARELIFE è stata somministrata a coloro che non avevano risposto alla terza wave. Agli intervistati le cui storie di vita erano già state raccolte nella wave 3 è stato sottoposto il regolare questionario SHARE (circa 13.000 intervistati).

L'intervista di SHARELIFE nella settima wave coinvolge diversi ambiti:

- i figli: domande retrospettive che riguardano il numero di figli, i figli deceduti, informazioni su gravidanze, nascite, adozioni, ecc.

- Il partner: domande retrospettive che rilevano le informazioni su tutte le relazioni avute fino al momento dell'intervista, quali convivenze, matrimoni, separazioni, divorzi ed eventuale morte del partner.

- L'alloggio: domande retrospettive sull'alloggio al momento dell'intervista e degli alloggi in passato.

- L'occupazione: domande retrospettive sui periodi di lavoro, comprese le informazioni sulla situazione lavorativa, le caratteristiche del lavoro, il reddito, il pensionamento benefici e l'occupazione dopo il pensionamento.

- la salute: domande retrospettive sulla salute durante l'infanzia e l'età adulta, compresi i dettagli su eventuali ricoveri in ospedale, malattie, infortuni, vaccinazioni, visite mediche, controlli preventivi, ecc.

- Sono state raccolte anche informazioni sulle circostanze dell'infanzia (ad esempio, salute dell'infanzia, rendimento scolastico, rapporto con i genitori, caratteristiche della sistemazione, libri letti, compagni).

- Per quanto riguarda le finanze (ad es. assicurazioni, alloggi, investimenti), sono state raccolte informazioni sugli investimenti finanziari che l'intervistato potrebbe aver fatto durante la sua vita.

- Sono state raccolte anche informazioni su eventi generali della vita (ad es. periodi di fame, periodi di felicità, stress, discriminazione sul lavoro, ecc.).

I questionari SHARE e SHARELIFE differiscono nella modalità di raccolta delle informazioni. L'intervista SHARELIFE viene somministrata completamente all'intervistato senza nessun coinvolgimento dei altri componenti della famiglia, a differenza di SHARE, in cui è possibile rispondere anche per gli altri familiari. Ad esempio, nel caso in cui una coppia intervistata abbia dei figli in comune, l'intervista SHARELIFE rileva la storia completa dei figli da entrambi i partner, mentre in SHARE è uno solo a rispondere alle domande sui figli (Bergmann et al., 2019).

1.4 Obiettivi

La fecondità in Europa è un aspetto in continua evoluzione nel corso degli anni. Inoltre, è stato riscontrato dalla letteratura che sono diversi i fattori che la influenzano.

Tra le cause potenziali, la crescita della partecipazione all'istruzione è spesso menzionata, ma le prove esplicite del suo ruolo come motore di questo cambiamento sono scarse (Ni Bhrolchain and Beaujouan, 2012).

A tal proposito, l'obiettivo della presente tesi è di indagare sul ruolo dell'istruzione nella fecondità. In particolare, la tesi sarà focalizzata sull'analisi del tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio in Europa.

Per l'obiettivo in questione, si analizzeranno i dati delle indagini SHARE e SHARELIFE, che riguardano 27 Paesi Europei, più Svizzera e Israele. L'indagine SHARE, progetto che aiuta a comprendere meglio le dinamiche di invecchiamento della popolazione, unita all'indagine SHARELIFE relativa al corso di vita degli intervistati, offre molteplici dati per provare a rispondere all'obiettivo del presente studio.

Il livello di istruzione, nei dati SHARE, è rilevato tramite due variabili: il numero di anni di istruzione e il livello di istruzione secondo la codifica

internazionale ISCED. Ai fini dell'analisi, si terrà conto del numero di anni di istruzione, in quanto più adatto allo studio della relazione in esame: da un lato permette di calcolare precisamente il numero di anni che trascorrono tra la fine dell'istruzione e la nascita del primo figlio; dall'altro lato può tener conto di eventuali riforme scolastiche (con conseguente incremento del numero di anni di scuola dell'obbligo).

Capitolo 2

I dati

2.1 Introduzione

Nel seguente capitolo vengono presentati i dati utilizzati per l'analisi. In particolare, viene descritta la prima fase di *pre-processing*, nonché le variabili ricostruite e la struttura del dataset ottenuto.

Inoltre, verranno presentate le principali analisi esplorative e descrittive dei dati oggetto di studio.

2.2 Pre-processing

Per raggiungere l'obiettivo del presente elaborato sono stati utilizzati i dati provenienti dall'indagine SHARE. In particolare, è stata utilizzata l'ultima release disponibile alla data di scrittura di questa tesi, che comprende, oltre alle precedenti, anche la wave 7 (Börsch-Supan et al., 2013), realizzata nel 2017.

In questa wave sono presenti due insiemi di dati: un dataset relativo all'indagine SHARE, e uno relativo all'indagine SHARELIFE, il quale contiene dati retrospettivi per i soggetti intervistati.

Le informazioni dei due dataset sono state incrociate tramite una variabile identificativa dell'intervistato, *mergeid*. Questa unione è stata effettuata per poter analizzare la relazione tra l'istruzione scolastica, rilevata

Anni di istruzione	0-5	6-8	9-11	12-13	14-16	17-21
Numerosità	8115	15371	17901	18655	12126	6630
%	10,29%	19,50%	22,71%	23,67%	15,38%	8,41%

Tabella 2.1: Numerosità classi di anni di istruzione

nell'indagine SHARE e le altre variabili retrospettive relative alla vita degli intervistati.

L'istruzione è stata rilevata tramite due differenti variabili: il numero di anni di istruzione e il livello di istruzione secondo il codice ISCED (UNESCO Institute for Statistics, 2012). Ai fini dell'analisi, è stato utilizzato il numero di anni di istruzione.

Gli anni di istruzione, per agevolare le analisi successive, sono stati trasformati in una variabile categoriale a sei classi. Il numero di classi è stato scelto sia per motivi di omogeneità di osservazioni nelle stesse, sia facendo riferimento al sistema scolastico italiano. Infatti, le classi previste sono 0-5 anni di istruzione, come sono gli anni della scuola primaria, 6-8, in riferimento alla scuola secondaria di primo grado, 9-11 i primi tre anni di scuola secondaria di secondo grado e 12-13 per gli ulteriori due anni, 14-16 si potrebbero riferire ad una laurea triennale e 17-21 ad una laurea magistrale. In Tabella 2.1 sono presenti il numero di osservazioni per ogni categoria di questa nuova variabile anni di istruzione.

Nel dataset SHARELIFE sono presenti, oltre ai dati demografici, alcune variabili socio-economiche, come, ad esempio, lo stato abitativo, *rsidstate*, rilevate dal quindicesimo all'ottantesimo anno di età dell'intervistato. Per ogni soggetto, quindi, si hanno sessantacinque variabili relative alla situazione abitativa, lavorativa ecc, ad ogni età del rispondente. Le variabili sono sessantaquattro per *hlthstate* che rileva i dati dal sedicesimo anno di età. Una descrizione più dettagliata di queste variabili è presente in Tabella 2.2.

Variabile	Descrizione	Tipo	Modalità
wrkstate	Stato lavorativo	Categoriale	1.Occupato full-time 2.Occupato part-time 3.Occupato indipendente 4.Disoccupato 5.Casalingo/a 6.Pensionato 7.Studente 8.Altro .a:Età non attribuibile
hwrkstate	Stato lavorativo (armonizzato)	Categoriale	1.Occupato dipendente 2.Occupato indipendente 3.Disoccupato 4.Casalingo/a 5.Pensionato 6.Studente 7.Altro .a:Età non attribuibile
prtnstate	Stato di convivenza con il partner	Categoriale	1.Vive solo 2.Vive con il partner .a:Età non attribuibile
chldstate	Numero totale di figli conviventi	Intero	
ychdstate	Numero totale di figli minorenni	Intero	
hlthstate	Stato di salute	Categoriale	1.Senza malattia 2.Con malattia 3.Malato per gran parte della vita .a:Età non attribuibile
rsidstate	Stato abitativo	Categoriale	1.Proprietario 2.Affittuario 3.All'estero 4.Casa dei genitori 5.Altro (es. usufrutto, etc.) .a:Età non attribuibile

Tabella 2.2: Variabili retrospettive di SHARELIFE.

2.2.1 Variabili ricostruite

A partire dai dati raccolti sono state create delle variabili di interesse per questa tesi.

In particolare, sono state ricostruite le seguenti variabili.

L'età al primo figlio, *eta_figlio*, è l'età in cui, in corrispondenza del numero di figli minori di 18 anni, variabile *ychdstate*, si trova il primo valore maggiore di zero.

L'età a fine istruzione, *eta_fine_istr* è l'età in cui compare per l'ultima volta "studente" nella variabile *workstate*. Tuttavia, quasi il 70% del campione non presenta mai la modalità "studente": probabilmente perchè la maggior parte di questi ha completato l'istruzione prima dei 15 anni e le informazioni dell'intervista vengono rilevate dopo i 15 anni di età -, e non è, quindi, possibile ricostruirla tramite la variabile *workstate*. Per questo motivo, per tutte le restanti osservazioni, è stato calcolato il dato come somma degli anni di istruzione, la variabile *raedyrs*, e un numero di anni, variabile per ogni Paese, relativo all'età in cui si inizia la scuola dell'obbligo. Per la maggior parte dei Paesi, l'istruzione obbligatoria inizia a sei anni, tranne per Bulgaria, Estonia, Croazia, Lettonia, Polonia, Lituania e Finlandia in cui comincia a sette e per Malta, a cinque anni (European Commission, 2018a).

Gli anni che intercorrono dalla fine dell'istruzione alla nascita del primo figlio, variabile *diff*, viene costruita come differenza tra l'età al primo figlio e l'età a fine istruzione. Da sottolineare che questa variabile può presentare anche valori minori di zero, quando il primo figlio nasce prima che la persona finisca il proprio percorso di studi.

Il numero di figli, *num_figli*, viene calcolata prendendo il valore massimo presente nella variabile numero di figli, *chldstate*.

Inoltre, sono state create altre tre variabili: lo stato lavorativo, lo stato di convivenza e lo stato abitativo all'età della nascita del primo figlio, ricostruite rispettivamente dalle variabili *hwrkstate*, *prtnstate* e *rsidstate*.

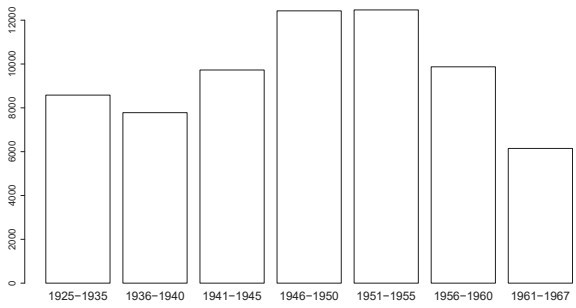


Figura 2.1: Distribuzione delle coorti di nascita.

2.2.2 Coorti di nascita

Nel dataset iniziale sono presenti informazioni relative a persone nate dal 1908 fino al 1990, quindi anche a persone più giovani del campione selezionato. Questo perchè, nonostante l'indagine SHARE faccia riferimento agli ultracinquantenni, sono idonee all'intervista la persona di ogni coppia selezionata per l'indagine di età pari o superiore a 50 anni e il suo partner, indipendentemente dall'età. Nell'analisi che verrà fatta in questa tesi verranno però considerate le persone con almeno 50 anni alla data dell'intervista.

Per le analisi, poi, è stata suddivisa la variabile anno di nascita, *rabyear*, in categoriale, ricreando delle coorti di nascita, cercando un compromesso tra omogeneità di osservazioni presenti in ciascuna e ampiezza delle stesse.

La divisione in coorti scelta, rappresentata nell'istogramma in Figura 2.1, è di sette classi, dal 1925 al 1935, dal 1936 al 1940, dal 1941 al 1945, dal 1946 al 1950, dal 1951 al 1955, dal 1956 al 1960 e dal 1961 al 1967. Le osservazioni relative a chi è nato dal 1908 al 1924 sono state eliminate, in quanto rappresentavano un campione ridotto per accorparle in una coorte a se stante e, allo stesso tempo, costruire una coorte di quasi trentanni dal 1908 al 1935 risultava eccessivo. Per questo motivo, è stato scelto di mantenere la prima coorte di dieci anni, 1925-1935, e di non aumentarla ulteriormente.

2.2.3 Il dataset finale

Il dataset di interesse per l'analisi risulta quindi composto da 78798 osservazioni e 13 variabili. Una descrizione più dettagliata delle variabili e della

Variabile	Descrizione	Tipo	Modalità
country	Paese di residenza	Categoriale	27 Paesi Europei, Svizzera e Israele
rabyear	Anno di nascita	Numerica	
ragender	Genere	Categoriale	1.Uomo 2.Donna
eta_figlio	Età al primo figlio	Numerica	
eta_fine_istr	Età a fine istruzione	Numerica	
diff	Differenza tra età al primo figlio e età a fine instruzione	Numerica	
stato_lav	Stato lavorativo al primo figlio	Categoriale	1.Occupato dipendente 2.Occupato indipendente 3.Disoccupato 4.Casalingo/a 5.Pensionato 6. Studente 7.Altro .a:Età non attribuibile
stato_part	Stato convivenza al primo figlio	Categoriale	1.Vive solo 2.Vive con il partner .a:Età non attribuibile
stato_abit	Stato abitativo al primo figlio	Categoriale	1.Proprietario 2.Affittuario 3.All'estero 4.Casa dei genitori 5.Altro (es. usufrutto, etc.) .a:Età non attribuibile
raedyrs	Anni di istruzione	Numerica	
raedyrs6	Anni di istruzione	Categoriale	0-5, 6-8, 9-11, 12-13, 14-16, 17-21
num_figli	Numero totale di figli	Numerica	
coorti	Coorti di nascita	Categoriale	1925-35, 1936-40, 1941-45, 1946-50, 1951-55, 1956-60, 1961-67.

Tabella 2.3: Variabili del dataset.

loro struttura è illustrata nella Tabella 2.3.

2.3 Analisi esplorative

Il campione analizzato comprende 78798 rispondenti, di cui il 56,73% sono donne.

Come già menzionato, sono state prese in considerazione sette coorti di nascita e sei categorie relative agli anni di istruzione dei rispondenti.

Il dataset contiene 29 Paesi: la numerosità campionaria per ogni Paese dell'indagine, in ordine crescente, è consultabile in Tabella 2.4.

La variabile risposta, chiamata *diff*, conta gli anni che intercorrono dalla fine dell'istruzione alla nascita del primo figlio.

Irlanda	**Cipro**	**Portogallo**	**Lussemburgo**	**Malta**	**Lettonia**
586	1049	1065	1104	1142	1393
0.74%	1.33%	1.35%	1.40%	1.44%	1.76%
Ungheria	**Israele**	**Finlandia**	**Bulgaria**	**Slovacchia**	**Lituania**
1404	1646	1678	1755	1773	1802
1.78%	2.08%	2.12%	2.22%	2.25%	2.28%
Romania	**Paesi Bassi**	**Croazia**	**Svizzera**	**Austria**	**Svezia**
1832	1980	2174	2490	3283	3363
2.32%	2.51%	2.75%	3.15%	4.16%	4.26%
Slovenia	**Danimarca**	**Grecia**	**Francia**	**Germania**	**Estonia**
3414	3438	3603	3712	4247	4518
4.33%	4.36%	4.57%	4.71%	5.38%	5.73%
Repubblica Ceca	**Spagna**	**Italia**	**Polonia**	**Belgio**	
4539	4630	4845	4895	5438	
5.76%	5.87%	6.14%	6.21%	6.90%	

Tabella 2.4: Osservazioni per paese nel dataset

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-25	4	8	8.70	12	54	9447

Tabella 2.5: Statistiche descrittive sulla variabile risposta.

Tale variabile può assumere valori negativi, quando la nascita del figlio avviene prima della fine dell'istruzione, oppure valori positivi, nel caso contrario. Nel campione osservato, il 2,5% delle osservazioni ha un figlio prima di completare il percorso scolastico.

Alcune statistiche descrittive relative alla variabile *diff* sono presenti in Tabella 2.5.

Un valore mancante, NA, sulla variabile risposta, invece, è dovuto ai casi in cui il rispondente non ha avuto nessun figlio. Nel campione utilizzato per le analisi, l'88,01% ha almeno un figlio, mentre il restante 11,98% non ha nessun figlio nel corso della propria vita.

Di particolare interesse è lo studio di differenze nelle biografie di queste due categorie di persone. Si è provato a rappresentare la sequenza della storia di queste persone.

In particolare, la Figura 2.2 mostra la differenza, tra individui senza figli e quelli con almeno un figlio, dello stato lavorativo nel corso del tempo. In corrispondenza di ogni età, dai 15 agli 80 anni, maggiore è il colore rappresentato, più è frequente quella modalità nello stato lavorativo. Non si notano differenze sostanziali, se non per una maggiore presenza di lavoro

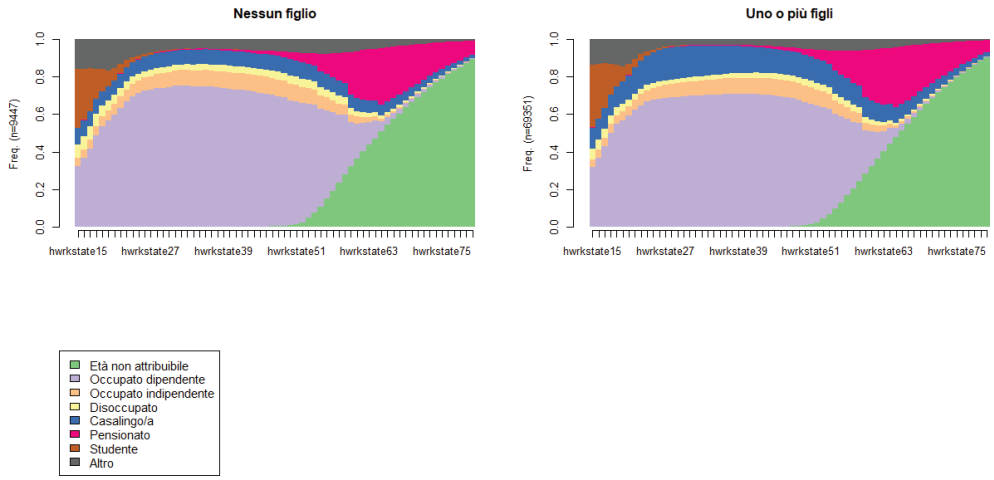


Figura 2.2: Distribuzione dello stato lavorativo distinto per chi ha e non ha figli

come casalingo/a per chi ha almeno un figlio.

Lo stato della partnership, come si può osservare nella Figura 2.3, differisce notevolmente nei due gruppi individuati. Oltre i 50 anni di età, le persone che non hanno avuto figli, per circa il 40% vive sola, a differenza di chi ha almeno un figlio, in cui a vivere solo è circa un 10% del campione.

Per quanto riguarda questi due grafici appena descritti, le sequenze partono dai 15 anni e arrivano agli 80 anni, ma non per tutti gli individui nel campione è stata effettivamente raggiunta l'età di 80 anni e, di conseguenza, i dati relativi alle età "non ancora raggiunte" non sono disponibili (nel grafico sono riportate in colore verde come "Età non attribuibile"). Ciò nonostante, per il confronto tra biografie di chi ha avuto figli e chi non ne ha avuto, questo non costituisce un problema, in quanto i dati "non completi" si hanno a partire dal cinquantesimo anno di età. Avere un figlio dopo i 50 anni è un evento possibile, ma non così altamente probabile e, quindi, anche se qualche individuo sperimentasse l'evento "primo figlio", non andrebbe ad alterare la situazione globale.

Da sottolineare, inoltre, che i due campioni osservati non hanno la stessa numerosità campionaria: le persone con figli sono 69351, mentre chi non ha

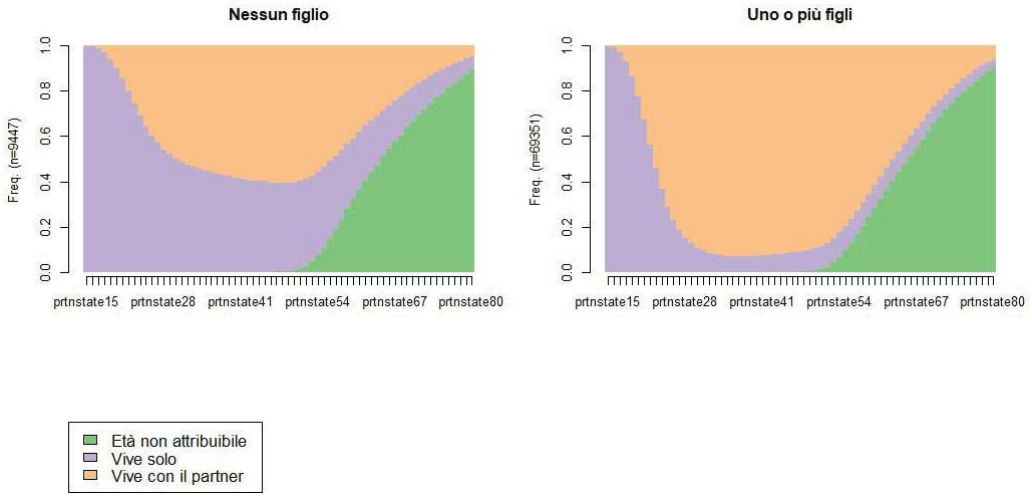


Figura 2.3: Distribuzione dello stato di convivenza distinto per chi ha e non ha figli

Anni di istruzione	Media del numero di figli
0-5	2.313617
6-8	2.271941
9-11	2.060276
12-13	1.936585
14-16	1.947221
17-21	1.898492

Tabella 2.6: Media del numero di figli totali per anni di istruzione.

figli sono meno di diecimila.

Ponendo l'attenzione solamente su chi ha avuto figli, nella Tabella 2.6 è descritta la variabile relativa al numero di figli totali, *num_figli*, in relazione agli anni di istruzione. Il valore più basso si riscontra per chi studia tra i 17 e i 21 anni, con una media di 1.89 figli; a seguire, chi studia 12 o 13 anni, con una media di 1.93 figli totali. Al contrario, il valore più alto si riscontra per chi studia al massimo per cinque anni, in cui la media di figli arriva a 2.31.

2.3.1 Confronto tra donne e uomini

In Figura 2.4 sono rappresentati gli anni che intercorrono dall'età fine dell'istruzione alla nascita del primo figlio, rispettivamente per uomini e donne e per numero di anni di istruzione.

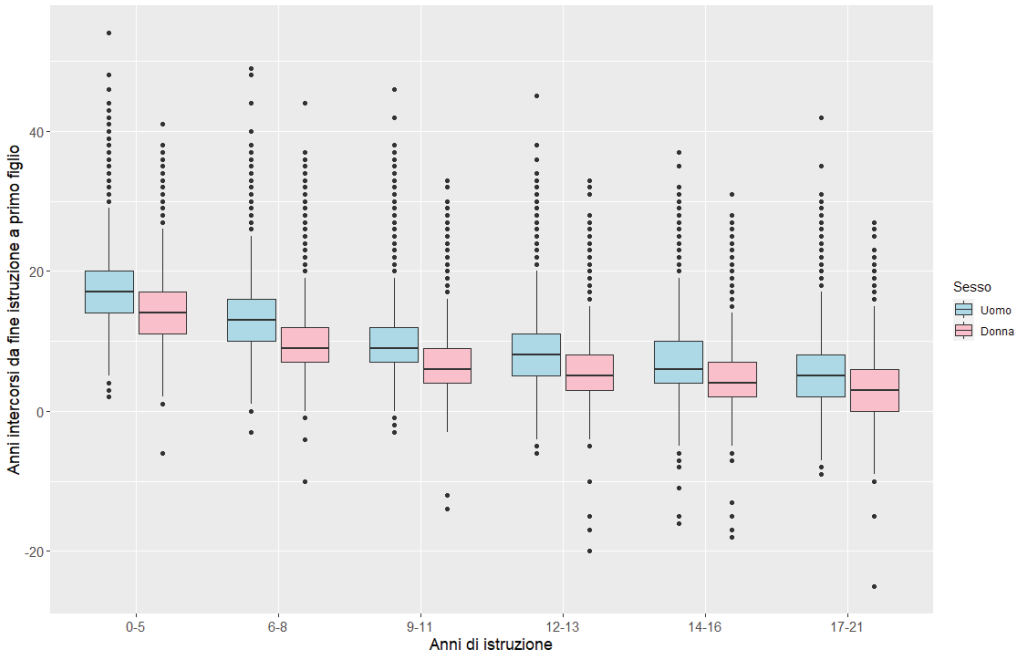


Figura 2.4: Anni che intercorrono tra istruzione e nascita primo figlio. Confronto tra donne e uomini.

Come è possibile notare, l'andamento è decrescente all'aumentare del numero di anni in cui la persona ha studiato. In particolare, chi trascorre al massimo cinque anni a scuola, aspetta, mediamente, almeno circa quindici anni prima di avere il primo figlio. Questo è abbastanza plausibile dal momento in cui queste persone finiscono l'istruzione, verosimilmente, all'età di 10, massimo 12 anni, e, di conseguenza, il tempo prima di poter avere un figlio aumenta obbligatoriamente per motivi riproduttivi. Dal lato opposto, per chi studia rientra nella categoria massima di anni di studio - e quindi, se consideriamo l'inizio della scuola a cinque anni, finisce di studiare almeno dopo i 22 anni di età - il tempo che intercorre per la nascita del primo figlio mediamente si avvicina ai cinque anni. Inoltre, il tempo che intercorre tra questi due eventi - la fine dell'istruzione e la nascita del primo figlio - è sempre mediamente minore per le donne rispetto agli uomini.

In Tabella 2.7 è rappresentata la distribuzione dello stato lavorativo alla nascita del primo figlio per uomini e donne.

Alla nascita del primo figlio, più del 90% degli uomini lavora come occupato dipendente o indipendente, mentre le donne sono poco più del 50%.

	Uomini	Donne
Occupato dipendente	0.819	0.509
Occupato indipendente	0.104	0.037
Disoccupato	0.012	0.037
Casalingo/a	0.004	0.348
Pensionato	0.002	0.002
Studente	0.009	0.009
Altro	0.047	0.054

Tabella 2.7: Stato lavorativo alla nascita del primo figlio. Confronto uomini e donne.

Una notevole differenza si può notare nel lavoro domestico o familiare: alla nascita del primo figlio, quasi il 35% delle donne ha un'occupazione di questo tipo, a differenza degli uomini che sono lo 0,4%. Nelle altre categorie non si riscontrano differenze marcate tra i due gruppi.

2.3.2 Paesi a confronto

Come si può osservare dalla Tabella 2.8, l'età al primo figlio varia da un'età media di 23.67 per la Bulgaria ai quasi 28 anni per l'Irlanda. L'Italia mediamente si avvicina ai 27 anni per l'età al primo figlio.

La Figura 2.5 evidenzia il tempo intercorso tra la fine istruzione e la nascita del primo figlio, per ogni Paese incluso nell'analisi SHARE. Portogallo, Italia e Spagna sembrano essere i Paesi in cui si aspetta di più per avere il primo figlio dopo aver finito di studiare (mediamente trascorrono più di dieci o undici anni). In Bulgaria, Repubblica Ceca ed Estonia, invece, gli anni che intercorrono sono mediamente cinque, i più bassi tra gli altri Paesi considerati.

Paese	Media età al primo figlio
Bulgaria	23.67526
Romania	24.12603
Repubblica Ceca	24.19627
Slovacchia	24.43083
Ungheria	24.43310
Croazia	24.65332
Slovenia	24.69034
Polonia	24.69586
Estonia	25.21443
Lettonia	25.24434
Lituania	25.26161
Austria	25.30266
Portogallo	25.43592
Cipro	25.68606
Israele	25.87149
Francia	25.88333
Germania	25.91076
Belgio	26.11583
Danimarca	26.15625
Finlandia	26.36935
Malta	26.45056
Svezia	26.48782
Paesi Bassi	26.68754
Spagna	26.88063
Italia	27.03648
Lussemburgo	27.40596
Grecia	27.43637
Svizzera	27.73413
Irlanda	27.92016

Tabella 2.8: Media dell'età al primo figlio per Paese, in ordine crescente.

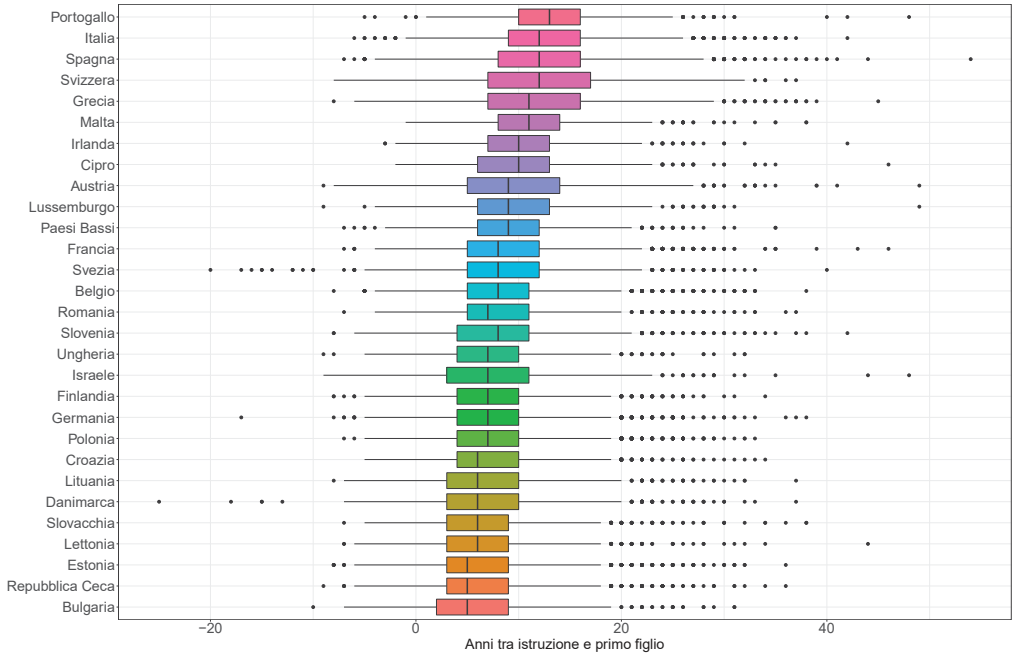


Figura 2.5: Anni che intercorrono tra istruzione e nascita primo figlio - Confronto tra Paesi.

Capitolo 3

Modelli per dati di sopravvivenza

3.1 Introduzione

In questo capitolo verranno introdotti, dal punto di vista teorico, i dati di durata e i relativi modelli per studiare la sopravvivenza. Per eventuali approfondimenti si faccia riferimento a Pace and Salvan (1996) e Klein and Moeschberger (2006).

Inoltre, verrà presentato il software R utilizzato per l'analisi dei dati e i principali pacchetti di rilevanza nell'ambito dei modelli per dati di durata.

3.2 Dati di durata

In molti settori è d'interesse analizzare i dati che rappresentano, per ciascuna unità statistica, il tempo trascorso dall'inizio dell'esperimento o dell'osservazione, fino al verificarsi di un evento.

L'analisi di dati di durata è frequente negli studi clinici, per il tempo di sopravvivenza, ad esempio, di un gruppo di pazienti trattati chirurgicamente; in ambito industriale, per ottenere informazioni sulla durata di corretto funzionamento di un dato prodotto o macchinario, ma anche nelle scienze sociali, economiche, demografiche, come ad esempio la durata di disoccupazione o di matrimonio.

Si consideri il tempo trascorso da un soggetto in un determinato stato, che si conclude con il verificarsi di uno specifico evento e ha inizio dal momento in cui il soggetto è esposto al rischio di subire tale evento. Il tempo all'evento, o tempo di sopravvivenza di un individuo, è la realizzazione di una variabile aleatoria continua T non negativa, con distribuzione tipicamente asimmetrica. La distribuzione di T è descritta dalle funzioni introdotte di seguito.

Supponendo che la variabile casuale T abbia una distribuzione di probabilità *F(t)* con *funzione di densità di probabilità f(t)*, la distribuzione di probabilità di T è data da:

$$F(t) = Pr(T < t) = \int_0^t f(u)\, du \tag{3.1}$$

e rappresenta la probabilità che il tempo di sopravvivenza sia inferiore ad un dato valore *t*.

La *funzione di sopravvivenza S(t)* esprime la probabilità di sperimentare l'evento dopo un certo istante t e può, quindi, essere utilizzata per definire la probabilità che un individuo sopravviva più di un certo tempo *t*.

La *funzione di sopravvivenza S(t)* è così definita:

$$S(t) = 1 - F(t) = Pr(T > t) = \int_t^\infty f(s)\, ds \tag{3.2}$$

S(t) è monotona decrescente: in corrispondenza di $t = 0$ nessun soggetto ha ancora sperimentato l'evento, quindi la probabilità di sopravvivenza è pari a 1; all'aumentare del tempo *t*, la funzione di sopravvivenza decresce in modo più o meno ripido a seconda della frequenza di sperimentazione dell'evento stesso.

La *funzione di rischio h(t)*, definita come:

$$h(t) = \lim_{\Delta t \to\ 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t))}{\Delta t} \tag{3.3}$$

può essere interpretata come la propensione al cambiamento di stato nell'istante *t*, per chi è nella condizione di farlo. La funzione *h(t)* è una funzione positiva e non è una probabilità; è utile per descrivere in che modo il rischio di sperimentare l'evento cambia nel tempo. La funzione di rischio,

infatti, può assumere un andamento crescente (l'evento tende a verificarsi alla fine del periodo di osservazione: tipico di unità soggette a invecchiamento o usura), decrescente (l'evento tende a verificarsi all'inizio del periodo di osservazione: casi in cui vi è una selezione iniziale), costante, ma anche non monotono.

Esiste una relazione tra la funzione di sopravvivenza *S(t)* e quella di rischio *h(t)*, data dalla formula seguente:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\ln(S(t))}{dt} \tag{3.4}$$

Un'altra quantità legata alle precedenti è la *funzione di rischio cumulato H(t)* definita da:

$$H(t) = \int_0^t h(s)\,ds = -\ln(S(t)), \tag{3.5}$$

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(s)\,ds\right), \tag{3.6}$$

che si può ricavare tramite la funzione di sopravvivenza *S(t)*, per avere un'idea del possibile andamento della funzione di rischio.

3.2.1 Schemi di osservazione

Nell'osservazione di tempi di sopravvivenza, lo schema di campionamento casuale semplice non è frequentemente adottato e si ricorre, invece, a schemi di osservazione con censura, di cui i principali sono detti censura del primo tipo, censura del secondo tipo, censura variabile, censura casuale.

Nella censura del primo tipo si mettono in osservazione, al tempo 0, n unità e si termina la prova dopo un tempo prefissato t_0 ($t_0 > 0$). Trascorso tale tempo, per alcune unità statistiche si sarà verificato l'evento d'interesse, e dunque sarà nota la durata di vita, mentre, per le rimanenti, si saprà solamente che la durata è maggiore di t_0.

Nella censura del secondo tipo si mettono in osservazione, al tempo 0, n unità statistiche e si termina la prova quando si è osservato l'evento di interesse per un numero prefissato di unità.

Lo schema di censura del primo tipo si generalizza facilmente allo schema di censura variabile, in cui il tempo di osservazione varia da unità ad unità. Questa situazione avviene, ad esempio, quando le unità entrano nella prova di durata a tempi diversi, ma lo studio deve essere completato entro un termine prefissato.

In altri casi, invece, il tempo di osservazione varia da soggetto a soggetto in modo non controllabile dallo sperimentatore. Un individuo può uscire dallo studio per vari motivi. In questi casi, è realistico assumere che i tempi di censura siano a loro volta realizzazioni di variabili casuali e si parlerà di censura casuale.

3.2.2 Stime della funzione di sopravvivenza

I metodi di analisi della sopravvivenza possono essere classificati in metodi non parametrici, metodi semi-parametrici e metodi parametrici, sulla base degli assunti che vengono fatti sulla distribuzione di T .

Uno dei metodi non parametrici per la stima della probabilità di sopravvivenza è il metodo del prodotto limite, noto anche come stimatore di Kaplan-Meier (Kaplan and Meier, 1958). Esso consiste nello stimare la probabilità condizionata di sopravvivenza in corrispondenza di ciascuno dei tempi in cui si verifica almeno un evento.

Siano $t_1 < t_2 < \ldots < t_D$ i tempi in cui si verificano uno o più eventi, e E_i il numero di eventi che accadono in t_i. Sia R_i il numero di soggetti a rischio di sperimentare l'evento al tempo t_i (cioè il numero di soggetti che non hanno ancora sperimentato l'evento in t_i o che sperimentano l'evento in t_i).

Lo stimatore di Kaplan-Meier è così definito:

$$\hat{S}_{KM}(t) = \begin{cases} 1, & \text{se } t < t_1 \\ \prod_{t_i \leq t}(1 - \frac{E_i}{R_i}), & \text{se } t_1 \leq \text{ t} \end{cases} \tag{3.7}$$

Lo stimatore ha valore costante dopo t_D e viene rappresentato graficamente da una curva a gradini continua a destra, con ordinata il tempo di

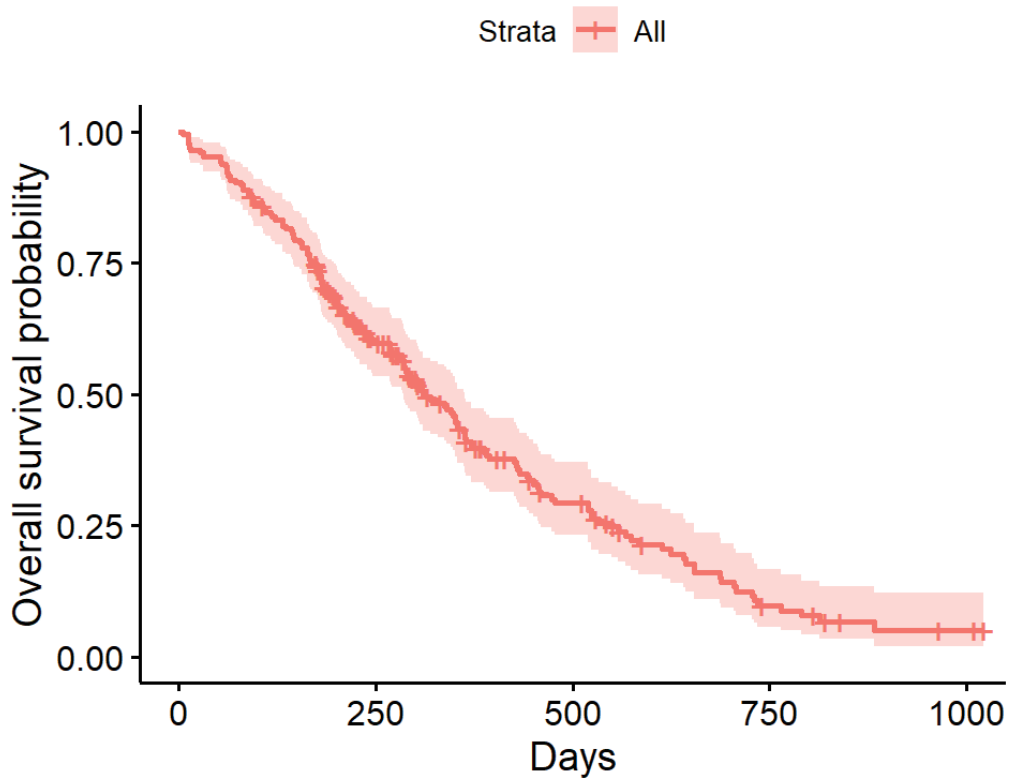


Figura 3.1: Esempio grafico di una curva di Kaplan-Meier.

sopravvivenza stimato e ascissa i tempi all'evento. Un esempio di curva di Kaplan-Meier è presente in Figura 3.1. La curva di Kaplan-Meier parte da 1 per $t = 0$ (nessun soggetto ha sperimentato l'evento), e decresce nel tempo: ha una caduta in ogni istante t_i in cui si verifica almeno un evento.

Lo stimatore di Kaplan-Meier risulta non distorto a varianza minima ed è caratterizzato da una distribuzione asintotica gaussiana con varianza stimabile tramite la formula di Greenwood (Pace and Salvan, 1996), che può essere utilizzata anche per il calcolo dell'intervallo di confidenza.

$$\hat{V}(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t)^2 \sum_{t_i \leq t} \frac{E_i}{R_i(R_i - E_i)} \tag{3.8}$$

Spesso, può essere utile confrontare stime della funzione di sopravvivenza calcolate per diversi *strati* di unità campionarie: ciascuno strato corrisponde tipicamente a differenti condizioni sperimentali, ad esempio trattati e non trattati. Un esempio grafico è presente in Figura 3.2. In questi casi, è possibile utilizzare lo stimatore di Kaplan-Meier per confrontare le curve di

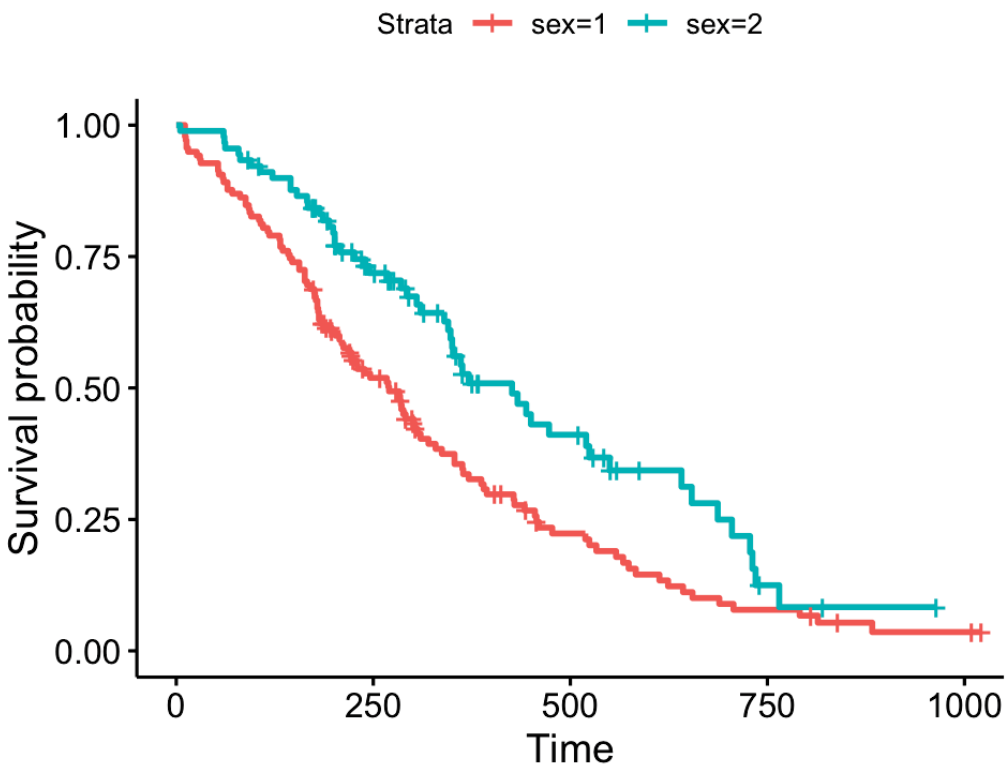


Figura 3.2: Esempio grafico di una curva di Kaplan-Meier stratificata per due gruppi.

sopravvivenza di due o più gruppi. Per valutare la differenza fra due o più curve di sopravvivenza si può ricorrere al *log-rank test*. La statistica log-rank test è costruita a partire da una serie ordinata di tabelle di contingenza 2×2, una per ognuno dei tempi dell'evento. Poiché, per ogni tabella, gli eventi si distribuiscono nei due gruppi proporzionalmente al numero di soggetti ancora a rischio, è possibile calcolare il numero di eventi attesi. Un eventuale scostamento, tra il totale degli eventi osservati e quelli attesi in ciascun gruppo, suggerisce un diverso livello di mortalità nei due gruppi (Mattiollo, 2012).

3.2.3 Modello di Cox

Nell'analisi di dati di sopravvivenza spesso è d'interesse lo studio della relazione fra il tempo all'evento e una serie di variabili esplicative. Il modello di Cox (1972) è un modello di regressione semiparametrico che esprime il

rischio in funzione del tempo e di alcune covariate. Dato un vettore di p covariate, $X^T = (X_1, ..., X_p)$ il modello si presenta nella forma

$$h(t|X) = h_0(t)exp(\beta^T X) = h_0(t)exp(\beta_1 X_1 + ... + \beta_p X_p), \qquad (3.9)$$

dove $\beta = (\beta_1, ..., \beta_p)^T$ è il vettore dei parametri e $h_0(t)$ è la funzione rischio di base. Quest'ultima è una componente non parametrica, positiva, ignota, uguale per tutti i soggetti, che dipende solo dal tempo e rappresenta la distribuzione della funzione di rischio per il gruppo di base ($X = 0$).

I coefficienti $\beta = (\beta_1, ..., \beta_p)^T$ misurano l'impatto delle covariate. Le quantità $exp(\beta_i)$ vengono chiamate *hazard ratios (HR)*. Un valore di β_i maggiore di 0, o, equivalentemente, un HR maggiore di 1, sta ad indicare che quando il valore della i-esima covariata aumenta, il rischio di sperimentare l'evento aumenta e, di conseguenza, la durata della sopravvivenza diminuisce.

Per riassumere,

- HR = 1: nessun effetto;
- HR < 1: riduzione del rischio;
- HR > 1: aumento del rischio.

L'ipotesi sottostante il modello è la proporzionalità dei rischi, ovvero il rapporto dei rischi di due individui con un diverso set di variabili esplicative è costante al variare del tempo. Si tratta di un modello robusto, in quanto i coefficienti di regressione approssimano bene i risultati di un corretto modello parametrico e fornisce informazioni primarie sulle funzioni di sopravvivenza e di rischio di diversi gruppi di individui con un numero minimo di assunzioni.

La caratteristica particolare di questo modello è permettere di fare inferenza senza richiedere necessariamente la specificazione di una classe parametrica per $h_0(t)$.

La stima dei parametri β del modello e le procedure di verifica d'ipotesi vengono effettuate tramite il metodo della massima verosimiglianza parziale. Questa produce stime non distorte e asintoticamente normali, e si comporta come una verosimiglianza propria sotto molti aspetti.

3.3 Software utilizzato per l'analisi

Per tutte le analisi è stato utilizzato R (R Core Team, 2017), un software open-source dedicato all'analisi statistica dei dati.

Per la parte di analisi esplorativa, in particolari per la rappresentazione grafica dei dati, sono stati utilizzati principalmente due pacchetti R, ggplot2 e TraMineR.

ggplot2 è un pacchetto per la produzione di grafici a partire da un set di dati. Diversamente dal pacchetto della distribuzione di base, però, ggplot2 prevede componenti separati che possono essere combinati in modi molto diversi.

TraMineR è un pacchetto per la visualizzazione e l'analisi di sequenze di dati categoriali. Il suo scopo principale è l'estrazione di conoscenza da sequenze di eventi o stati che descrivono il corsi della vita, sebbene la maggior parte delle sue caratteristiche si applicano anche a dati non temporali, come ad esempio sequenze di testo o DNA. Il nome TraMineR deriva da Life Trajectory Miner for R. Nel secondo capitolo di questa tesi, sono stati in particolare utilizzati i comandi *seqdef*, per la creazione delle sequenze e *seqdplot* per generare i grafici dell'andamento delle sequenze nel tempo (Figure 2.2 e 2.3 del secondo capitolo).

Per la parte di modellazione statistica, sono state utilizzate le seguenti librerie: survival, per l'analisi di sopravvivenza e survminer, che fornisce delle funzioni per la visualizzazione dei risultati dell'analisi di sopravvivenza, in particolare tramite il comando *ggsurvplot*. La funzione *coxph()* del pacchetto survival può essere utilizzata per generare il modello di Cox a rischi proporzionali.

Per la diagnostica del modello di Cox è stata utilizzata la funzione *cox.zph()* e per il log-rank test la funzione *survdiff()*.

Capitolo 4

I risultati

4.1 Introduzione

Nel quarto capitolo vengono presentate le applicazioni dei modelli ai dati.

Si procederà analizzando "dal generale al particolare": inizialmente, verrà introdotto un modello *baseline* con tutte le esplicative d'interesse, per poi procedere approfondendo alcuni aspetti rilevanti. In particolare, si farà una prima presentazione sugli anni di istruzione e, a seguire, si discuteranno le coorti di nascita, il genere dei rispondenti e l'area geografica di provenienza.

Infine, si conclude con una breve discussione sui risultati ottenuti.

4.2 Preparazione dei dati

Obiettivo di questo studio è l'analisi di vari aspetti inerenti il tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio in vari paesi Europei. Le ipotesi teoriche di partenza suggeriscono come l'istruzione giochi un ruolo importante nelle tempistiche della prima gravidanza, così come altri fattori sociali ed economici evidenziati nel primo capitolo.

L'interesse su cui si focalizza questo elaborato non è tanto l'età alla prima maternità - argomento ampiamente discusso in letteratura -, ma la durata tra i due eventi appena menzionati: la fine dell'istruzione e la nascita del

Occupato dipendente	Occupato indipendente	Disoccupato	Casalingo/a	Altro
43476	4511	1811	13685	3518
64,88%	6,73%	2,70%	20,42%	5,25%

Tabella 4.1: Numero di osservazioni della variabile *stato_ lav*.

primo figlio. In particolare, si vuole comprendere se esiste una accelerazione o una decelerazione di questo tempo in base agli anni di studio effettuati.

Il dataset utilizzato per l'analisi proviene dall'indagine SHARELIFE (Börsch-Supan et al., 2013). Una descrizione dettagliata del campione di dati con le relative variabili d'interesse è riportato nel secondo capitolo.

Per l'analisi e l'implementazione dei modelli sono stati apportati alcuni accorgimenti in aggiunta a quanto già presentato nel capitolo secondo.

In primo luogo, la variabile *stato_ lav*, relativa allo stato lavorativo alla nascita del primo figlio, è stata ricodificata come segue:

- sono state eliminate le osservazioni di chi aveva "full-time education" come stato lavorativo, in quanto non rappresentano un lavoro e gli anni di istruzione sono già conteggiati nella variabile che conta il numero di anni di istruzione divisa per classi, *raedyrs6*.

- Le osservazioni relative ai pensionati sono state accorpate nella categoria *Altro*, per motivi di numerosità campionaria.

La variabile *stato_ lav* ora appare composta come in Tabella 4.1

Inoltre, sono state eliminate le osservazioni relative a chi ha avuto un figlio prima di finire l'istruzione. Nel campione osservato, il 2,5% delle osservazioni ha un figlio prima di completare il percorso scolastico. Nonostante sarebbe un aspetto interessante da analizzare, è stata presa questa scelta in quanto lo studio è incentrato sulla valutazione del tempo intercorso dopo la fine dell'istruzione, e non prima.

Per una migliore gestione dei Paesi nelle analisi successive, a partire dalla variabile *country* è stata creata una nuova variabile categoriale a quattro classi che indica, appunto, l'area geografica del Paese di riferimento: Nord-Europa, Centro-Europa, Sud-Europa ed Est-Europa. Nella Tabella 4.2, so-

Categoria	**Nord-Europa**	**Centro-Europa**	**Sud-Europa**	**Est-Europa**
Paesi	Danimarca Estonia Finlandia Svezia Lettonia Lituania Irlanda	Germania Svizzera Austria Paesi Bassi Francia Belgio Lussemburgo	Spagna Portogallo Italia Grecia Malta Cipro Israele	Polonia Ungheria Slovacchia Repubblica Ceca Romania Bulgaria Slovenia Croazia
N° Oss.	13985	18632	15174	19210
%	20,87%	27,80%	22,64%	28,67%

Tabella 4.2: Paesi e numero di osservazioni per ogni categoria di *area*.

no indicati i Paesi corrispondenti ad ogni categoria e il relativo numero di osservazioni risultanti da questa suddivisione.

Da sottolineare che Israele è stato inserito nella categoria Sud-Europa, anche se non propriamente corretto, in quanto non rientra tra i Paesi Europei. Si potrebbe valutare in futuro di considerarlo singolarmente e confrontarlo con un Paese per ognuna delle quattro categorie europee, in modo da evidenziare eventuali similitudini o differenze con queste ultime.

Il dataset finale, utilizzato nelle successive analisi, è ora composto da 67001 osservazioni e 14 variabili. La variabile risposta è *diff*, e conta, per ogni soggetto, il numero di anni che sono passati dalla fine dell'istruzione alla nascita del primo figlio.

Nei paragrafi successivi verranno spesso menzionati i termini *evento* e *rischio*. L'*evento* d'interesse di questo studio è la nascita del primo figlio e il tempo per sperimentare tale evento è calcolato a partire dalla fine del percorso scolastico. Per *rischio*, invece, si intende la probabilità di sperimentare la nascita del primo figlio. Più precisamente, dati due gruppi di persone A e B, se nel gruppo A il rischio di sperimentare l'evento è maggiore di quello del gruppo B, significa che il gruppo A è più propenso a ridurre il tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio.

4.3 Modello *baseline*

In primo luogo, è stato analizzato un modello generale per valutare l'ipotesi che gli anni di istruzione rallentino o accelerino il tempo che intercorre tra i due eventi "fine istruzione" e "primo figlio".

A tale scopo, è stato creato un primo modello *baseline*, un modello di Cox che prevede come covariate, in primis, il numero di anni di istruzione e, a seguire, le coorti di nascita, il genere, l'area geografica e lo stato lavorativo alla nascita del primo figlio.

L'output del modello stimato tramite il software R è presentato in Tabella 4.3.

Da questo primo modello generale si può osservare come tutti i coefficienti siano altamente significativi, tranne il coefficiente relativo alla categoria *employed* dello stato lavorativo che non risulta statisticamente significativo.

Per la variabile anni di istruzione, in cui la categoria di riferimento è 0-5 anni, i coefficienti sono tutti positivi: l'appartenenza ad ognuna delle categorie di anni istruzione, fa aumentare il rischio di sperimentare l'evento (cioè la nascita del primo figlio) prima rispetto alla categoria di riferimento 0-5 anni. Inoltre, tale rischio cresce sempre di più con gli anni passati a scuola.

Il medesimo commento si può fare per le coorti di nascita e per il genere: appartenere ad una coorte più "giovane" o essere donna aumenta il rischio di sperimentare l'evento. In altre parole, essere donna oppure appartenere ad una coorte diversa da quella di riferimento (1925-1935) porta ad accorciare il tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio.

Al contrario, i coefficienti relativi alle variabili area geografica e stato lavorativo al primo figlio sono negativi. In questo caso, l'appartenenza alle relative classi diminuisce il rischio di ridurre il tempo tra i due eventi "fine istruzione" e "nascita del primo figlio", rispetto alla categoria di riferimento.

Tutte queste valutazioni sono da considerarsi al netto delle altre covariate.

Coefficienti	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**	
Anni di istruzione (*rif. 0-5*)						
6-8	0.498382	1.646056	0.014865	33.528	< 2e-16	***
9-11	0.903447	2.468097	0.015020	60.150	< 2e-16	***
12-13	1.141016	3.129948	0.015038	75.878	< 2e-16	***
14-16	1.309417	3.704012	0.016596	78.901	< 2e-16	***
17-21	1.479947	4.392714	0.019606	75.483	< 2e-16	***
Coorte (*rif. 1925-1935*)						
1936-1940	0.165340	1.179794	0.015701	10.530	< 2e-16	***
1941-1945	0.221192	1.247562	0.014946	14.799	< 2e-16	***
1946-1950	0.301468	1.351842	0.014292	21.094	< 2e-16	***
1951-1955	0.320367	1.377634	0.014393	22.258	< 2e-16	***
1956-1960	0.329886	1.390810	0.015172	21.743	< 2e-16	***
1961-1967	0.324106	1.382794	0.017210	18.832	< 2e-16	***
Genere (*rif. Uomini*)						
Donne	0.444425	1.559593	0.008889	49.997	< 2e-16	***
Area geografica (*rif. Est-Europa*)						
Nord-Europa	-0.240485	0.786247	0.011379	-21.133	< 2e-16	***
Centro-Europa	-0.422314	0.655528	0.010605	-39.822	< 2e-16	***
Sud-Europa	-0.573564	0.563513	0.011358	-50.499	< 2e-16	***
Stato lavorativo (*rif. Disoccupato*)						
Occupato dipendente	-0.039651	0.961125	0.024258	-1.635	0.10215	
Occupato indipendente	-0.142456	0.867226	0.028118	-5.066	4.06e-07	***
Casalingo/a	0.079911	1.083191	0.025250	3.165	0.00155	**
Altro	0.194420	1.214606	0.029030	6.697	2.12e-11	***

Tabella 4.3: Output del modello *baseline*.

	chisq	**df**	**p**
Anni di istruzione	6168	5	<2e-16
Coorti	1604	6	<2e-16
Genere	498	1	<2e-16
Area geografica	1796	3	<2e-16
Stato lavorativo	374	4	<2e-16
GLOBAL	8884	19	<2e-16

Tabella 4.4: Diagnostica del modello *baseline*.

Per definizione (Klein and Moeschberger, 2006), il modello di Cox a rischi proporzionali ipotizza che i coefficienti non dipendano dal tempo t. Per validare il modello stimato, tale ipotesi deve essere verificata.

Tramite la funzione *cox.zph* di R è stata testata l'ipotesi per ognuna delle variabili del modello e per il modello nel suo complesso. I risultati ottenuti sono presentati in Tabella 4.4 e purtroppo l'ipotesi di rischio proporzionale viene rigettata.

4.3.1 Anni di istruzione

Per analizzare il ruolo degli anni di istruzione nel tempo che intercorre tra il percorso scolastico e la prima gravidanza è stata rappresentata la curva

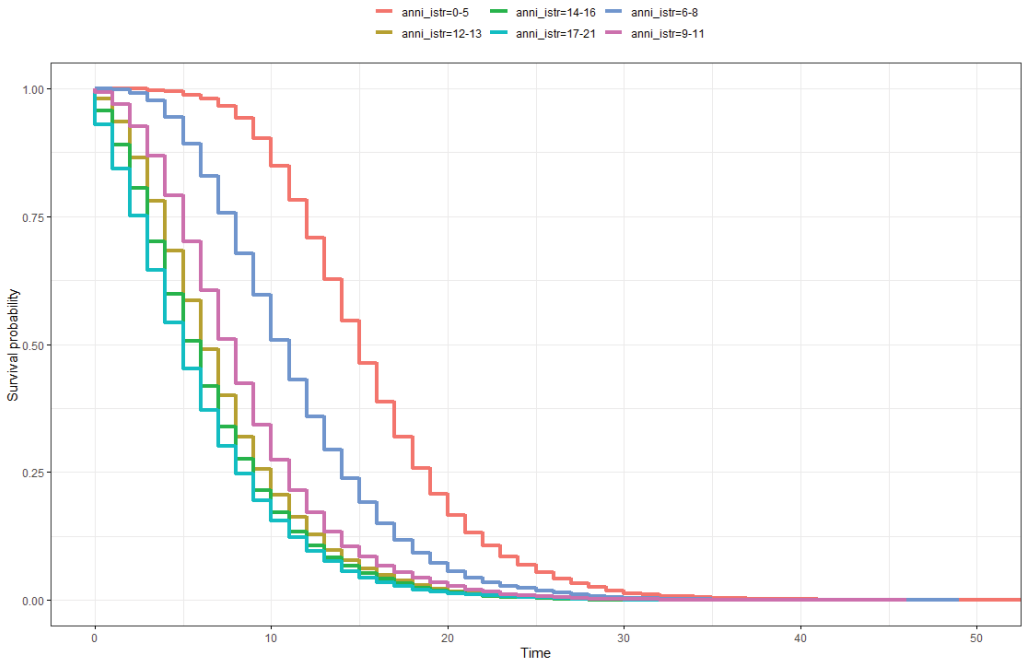


Figura 4.1: Curva di Kaplan-Meier per anni di istruzione.

di sopravvivenza di Kaplan-Meier, per ogni categoria di anni di istruzione effettuati.

Nella Figura 4.1, è possibile notare che il rischio di ridurre il tempo per la nascita del primo figlio aumenta per chi ha studiato più anni: chi ha studiato per almeno 17-21 anni è più propenso, rispetto a chi ha studiato per meno anni, ad accorciare il tempo per la nascita del primo figlio.

Per confrontare le curve di sopravvivenza, è stato inoltre eseguito il *log-rank test*, i cui risultati sono riportati in Tabella 4.5. Il p-value è prossimo allo zero, quindi non si accetta l'ipotesi nulla: le curve di sopravvivenza stimate in base agli anni di istruzione sono significativamente diverse tra loro.

Il modello di Cox univariato con variabile risposta *diff* e come esplicativa solamente il numero di anni di istruzione, ha prodotto i risultati presenti nella Tabella 4.6.

I coefficienti del modello sono tutti statisticamente significativi e crescono con gli anni di studio, in linea con quanto visto nella curva di Kaplan-Meier. La diagnostica del modello, inoltre, è presente in Tabella 4.7: l'ipotesi di rischi proporzionali anche in questo caso viene purtroppo respinta.

Una ulteriore visione di quanto è stato ottenuto, si può osservare in Tabel-

	N	**Observed**	**Expected**	$(O-E)^2/E$	$(O-E)^2/V$
0-5	7105	7105	14777	3983	6340
6-8	13542	13542	17904	1063	1683
9-11	15693	15693	13766	270	391
12-13	16209	16209	11683	1754	2450
14-16	9743	9743	6157	2089	2630
17-21	4709	4709	2714	1466	1735
Chisq= 13153 on 5 degrees of freedom, p= <2e-16					

Tabella 4.5: *Log-rank test* - Confronto curve di sopravvivenza per anni di istruzione.

	coef	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**	
(*rif. 0-5*)						
6-8	0.51922	1.68072	0.01473	35.25	<2e-16	***
9-11	0.98225	2.67047	0.01451	67.67	<2e-16	***
12-13	1.19950	3.31847	0.01450	82.70	<2e-16	***
14-16	1.34160	3.82514	0.01588	84.48	<2e-16	***
17-21	1.43796	4.21211	0.01903	75.56	<2e-16	***

Tabella 4.6: Output del modello di Cox univariato con epslicativa gli anni di istruzione.

	chisq	**df**	**p**
Anni di istruzione	5830	5	<2e-16
GLOBAL	5830	5	<2e-16

Tabella 4.7: Diagnostica del modello di Cox univariato con epslicativa gli anni di istruzione.

Anni di istruzione	Media *diff*
0-5	15.70
6-8	11.37
9-11	8.42
12-13	7.24
14-16	6.50
17-21	5.99

Tabella 4.8: Media della variabile *diff* per anni di istruzione.

la 4.8. Con l'aumentare degli anni di istruzione effettuati dagli intervistati, la media di anni che passano dalla fine dell'istruzione alla nascita del primo figlio diminuisce sempre di più.

Per riassumere, da queste prime analisi si potrebbe ipotizzare che chi studia per un numero maggiore di anni tende a far diminuire gli anni per la nascita del primo figlio, dopo la fine del proprio percorso scolastico.

4.3.2 Coorti di nascita

Ciò nonostante, in queste analisi non è stato approfondito un aspetto di notevole rilevanza: con il passare degli anni, il livello di istruzione in Europa è aumentato. Gli anni di istruzione sono, quindi, aumentati sempre di più con le coorti di nascita.

Queste differenze risultanti dalle analisi sul numero di anni di istruzione potrebbero, perciò, essere dovute alle coorti di nascita: il livello di istruzione aumenta, di conseguenza si finisce di studiare ad un'età più elevata e gli individui potrebbero essere "costretti" ad anticipare la nascita del primo figlio per motivi di fertilità.

Sono state analizzate, quindi, le coorti di nascita.

Nella Figura 4.2 sono presenti le curve di sopravvivenza di Kaplan-Meier stimate per ogni coorte di nascita. L'andamento è decrescente: man mano che le coorti avanzano, quindi si procede verso individui più giovani, il rischio di ridurre il tempo per la nascita del primo figlio aumenta sempre di più.

Per verificare quanto visto nelle curve di sopravvivenza per coorti di nascita, è stato eseguito il *log-rank test*, i cui risultati sono riportati in Tabella 4.9.

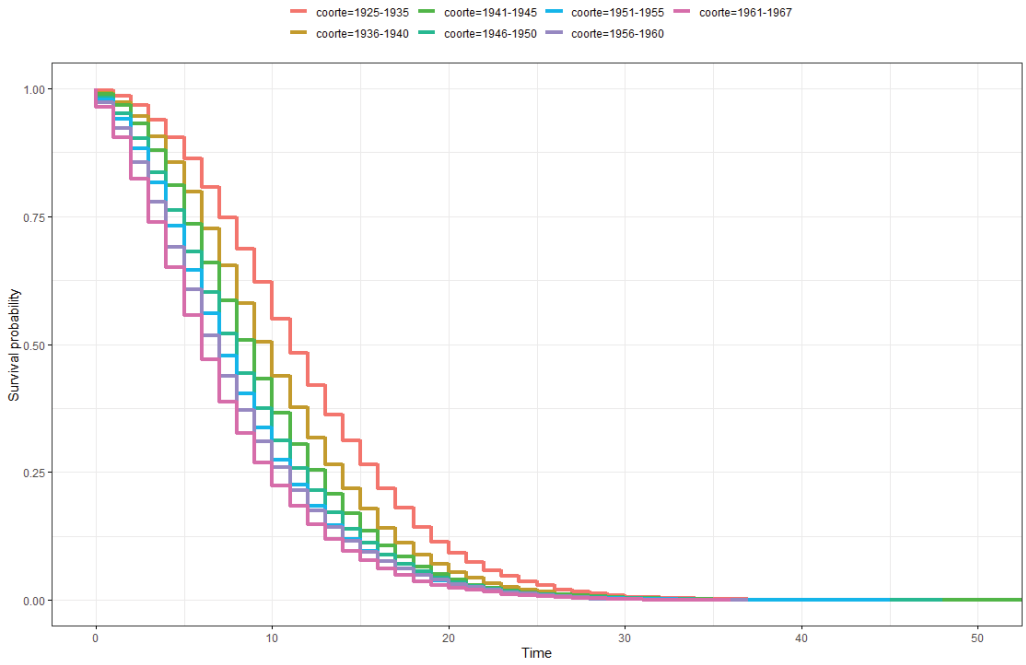


Figura 4.2: Curva di Kaplan-Meier per coorti di nascita.

Il p-value è prossimo allo zero, quindi rifiutiamo l'ipotesi nulla: l'andamento della sopravvivenza nei gruppi - coorti di nascita - è significativamente diverso.

A conferma di quanto è evidenziato dalla curva di sopravvivenza, è stato implementato un modello di Cox per ogni coorte diversa, utilizzando come unica covariata gli anni di istruzione.

Nella Tabella 4.10 sono riportati i coefficienti risultanti dai modelli stimati per ognuna delle sette coorti di nascita.

	N	**Observed**	**Expected**	$(O-E)^2/E$	$(O-E)^2/V$
1925-1935	8581	8581	12506	1231.7	1782.5
1936-1940	7780	7780	9265	238.0	319.4
1941-1945	9727	9727	10161	18.5	25.1
1946-1950	12426	12426	11645	52.4	72.8
1951-1955	12468	12468	10766	269.1	367.7
1956-1960	9872	9872	8118	378.8	493.9
1961-1967	6147	6147	4540	568.7	696.7
Chisq= 3234 on 6 degrees of freedom, p= <2e-16					

Tabella 4.9: *Log-rank test* - Confronto curve di sopravvivenza per coorti.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Coorte 1925-1935						
Anni di istruzione (*rif. 0-5*)						
6-8	0.54508	1.72474	0.03034	17.97	<2e-16	***
9-11	0.99143	2.69507	0.03496	28.36	<2e-16	***
12-13	1.16041	3.19123	0.03725	31.15	<2e-16	***
14-16	1.42647	4.16399	0.04305	33.14	<2e-16	***
17-21	1.59379	4.92237	0.05665	28.13	<2e-16	***
Coorte 1936-1940						
Anni di istruzione (*rif. 0-5*)						
6-8	0.51216	1.66888	0.03565	14.36	<2e-16	***
9-11	0.95971	2.61095	0.03857	24.88	<2e-16	***
12-13	1.15203	3.16460	0.03971	29.01	<2e-16	***
14-16	1.43676	4.20705	0.04487	32.02	<2e-16	***
17-21	1.51689	4.55801	0.05507	27.54	<2e-16	***
Coorte 1941-1945						
Anni di istruzione (*rif. 0-5*)						
6-8	0.50826	1.66240	0.03645	13.94	<2e-16	***
9-11	0.94086	2.56217	0.03701	25.42	<2e-16	***
12-13	1.11934	3.06282	0.03787	29.56	<2e-16	***
14-16	1.29731	3.65944	0.04105	31.60	<2e-16	***
17-21	1.43908	4.21680	0.05111	28.16	<2e-16	***
Coorte 1946-1950						
Anni di istruzione (*rif. 0-5*)						
6-8	0.50434	1.65589	0.03611	13.97	<2e-16	***
9-11	0.90834	2.48020	0.03493	26.01	<2e-16	***
12-13	1.16595	3.20898	0.03508	33.23	<2e-16	***
14-16	1.31805	3.73614	0.03818	34.52	<2e-16	***
17-21	1.38651	4.00084	0.04505	30.78	<2e-16	***
Coorte 1951-1955						
Anni di istruzione (*rif. 0-5*)						
6-8	0.52149	1.68454	0.04108	12.70	<2e-16	***
9-11	0.92726	2.52757	0.03923	23.63	<2e-16	***
12-13	1.10722	3.02593	0.03889	28.47	<2e-16	***
14-16	1.21129	3.35782	0.04127	29.35	<2e-16	***
17-21	1.32648	3.76775	0.04757	27.88	<2e-16	***
Coorte 1956-1960						
Anni di istruzione (*rif. 0-5*)						
6-8	0.59859	1.81955	0.04958	12.07	<2e-16	***
9-11	0.95567	2.60042	0.04535	21.07	<2e-16	***
12-13	1.19574	3.30601	0.04508	26.52	<2e-16	***
14-16	1.28364	3.60974	0.04757	26.98	<2e-16	***
17-21	1.32319	3.75539	0.05400	24.50	<2e-16	***
Coorte 1961-1967						
Anni di istruzione (*rif. 0-5*)						
6-8	0.51649	1.67613	0.07414	6.966	3.25e-12	***
9-11	0.90216	2.46493	0.06678	13.510	< 2e-16	***
12-13	1.02842	2.79664	0.06553	15.694	< 2e-16	***
14-16	1.09291	2.98295	0.06905	15.828	< 2e-16	***
17-21	1.20358	3.33204	0.07441	16.176	< 2e-16	***

Tabella 4.10: Output dei modelli stimati per ogni coorte.

	chisq	df	p
Coorte 1925-1935			
Anni di istruzione	914	5	<2e-16
GLOBAL	914	5	<2e-16
Coorte 1936-1940			
Anni di istruzione	792	5	<2e-16
GLOBAL	792	5	<2e-16
Coorte 1941-1945			
Anni di istruzione	858	5	<2e-16
GLOBAL	858	5	<2e-16
Coorte 1946-1950			
Anni di istruzione	1086	5	<2e-16
GLOBAL	1086	5	<2e-16
Coorte 1951-1955			
Anni di istruzione	908	5	<2e-16
GLOBAL	908	5	<2e-16
Coorte 1956-1960			
Anni di istruzione	555	5	<2e-16
GLOBAL	555	5	<2e-16
Coorte 1961-1967			
Anni di istruzione	214	5	<2e-16
GLOBAL	214	5	<2e-16

Tabella 4.11: Diagnostica dei modelli stimati per ogni coorte.

Innanzitutto, i coefficienti sono tutti altamente significativi. In secondo luogo, si può notare come le stime dei coefficienti siano molto simili - per la stessa categoria di anni di istruzione - nelle varie coorti considerate. Ad esempio, il coefficiente stimato di 6-8 di anni di istruzione della prima coorte è molto simile ai coefficienti di 6-8 di istruzione nelle altre coorti.

La diagnostica dei modelli, inoltre, è consultabile in Tabella 4.11: per tutti i modelli stimati per ogni coorte, l'ipotesi di rischi proporzionali viene rigettata.

Come già menzionato, in letteratura è ampiamente sostenuto che il livello di istruzione si sia modificato nel corso degli anni, in particolare aumentando. Per il campione a disposizione gli anni di nascita considerati vanno dal 1925 al 1967 e si è voluto indagare se tale teoria è valida anche per questo campione di dati.

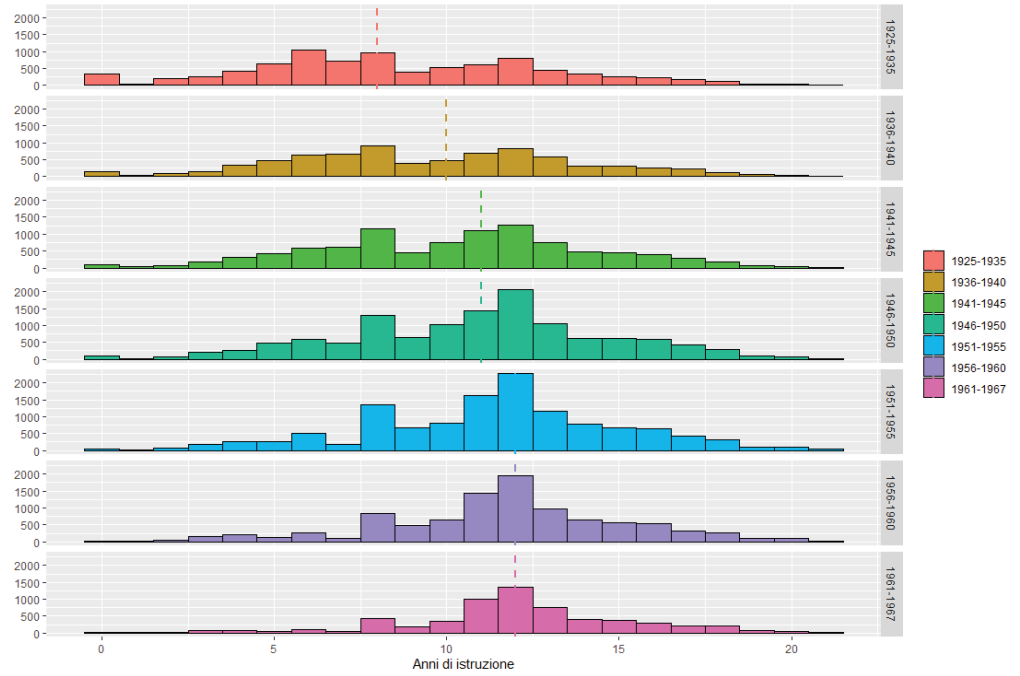


Figura 4.3: Distribuzione degli anni di istruzione per ogni coorte di nascita.

Nella Figura 4.3 sono visualizzate le distribuzioni del numero di anni di istruzione, per ogni coorte osservata e con una linea tratteggiata ad indicare la mediana della distribuzione: si nota chiaramente come il numero di anni passati sui banchi di scuola sia aumentato con il passare delle coorti di nascita. La mediana è stata scelta in quanto più adeguata in presenza di *valori anomali*.

Inoltre, dato che il livello di istruzione è aumentato negli anni, si è voluto indagare se anche l'età alla nascita del primo figlio viene traslata oppure no.

Nella Figura 4.4 sono raffigurate le distribuzioni dell'età al primo figlio, per ogni coorte osservata nel campione e con una linea tratteggiata ad indicare la mediana della distribuzione. Anche in questo caso è scelta la mediana perchè le distribuzioni dell'età al primo figlio per ogni coorte presentano delle code a destra molto marcate.

In questo caso, la prima coorte si distacca dalle altre in termini di mediana, questo probabilmente dovuto sia ad un motivo di numerosità campionaria, che di ampiezza della coorte, in quanto è formata da dieci anni e non da cinque come nelle altre. Nelle altre coorti, invece, non ci sono differenze sostanziali nell'età alla nascita del primo figlio.

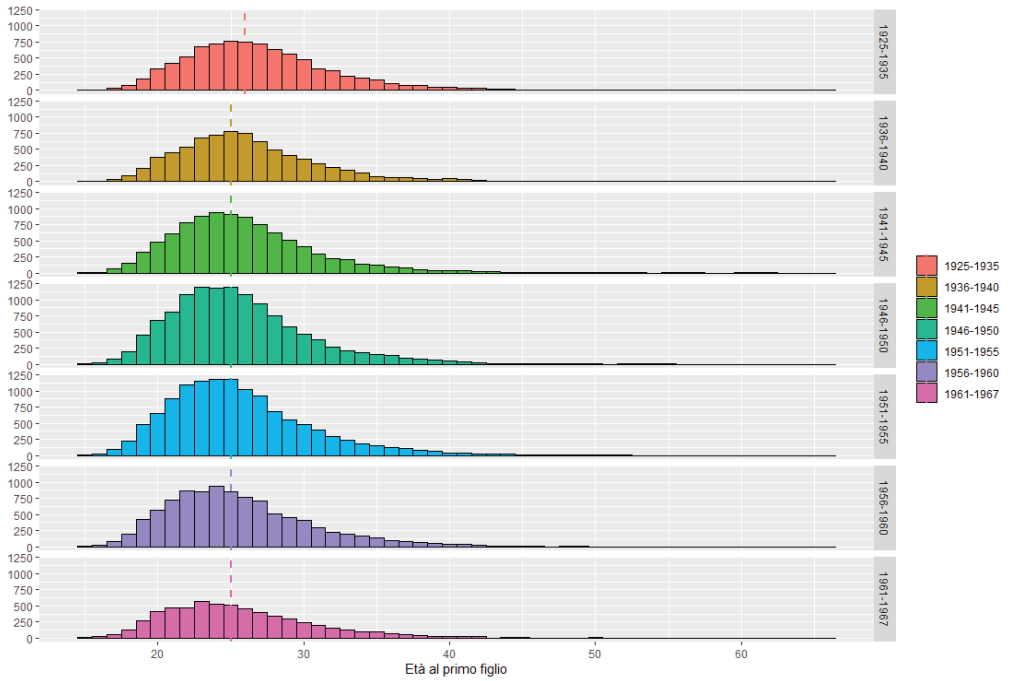


Figura 4.4: Età al primo figlio per coorte di nascita.

In questo campione, con quanto già sottolineato in merito alla prima coorte, il titolo di studio non ha un effetto di accelerazione o decelerazione del tempo che intercorre tra i due eventi. Nonostante, quindi, aumentino i livelli di istruzione negli anni, l'età al primo figlio rimane invariata.

Donne e uomini

Nella Figura 4.5 sono invece rappresentate le curve di sopravvivenza di Kaplan-Meier divise per genere.

Come ci si poteva aspettare per motivi di fertilità, il rischio di sperimentare prima l'evento "primo figlio" è maggiore per le donne, rispetto agli uomini.

In Tabella 4.12 è riportato il *log-rank test*, per confrontare le curve di sopravvivenza in base al genere. Il p-value porta a rifiutare l'ipotesi nulla: l'andamento della sopravvivenza per le donne egli uomini è significativamente diverso.

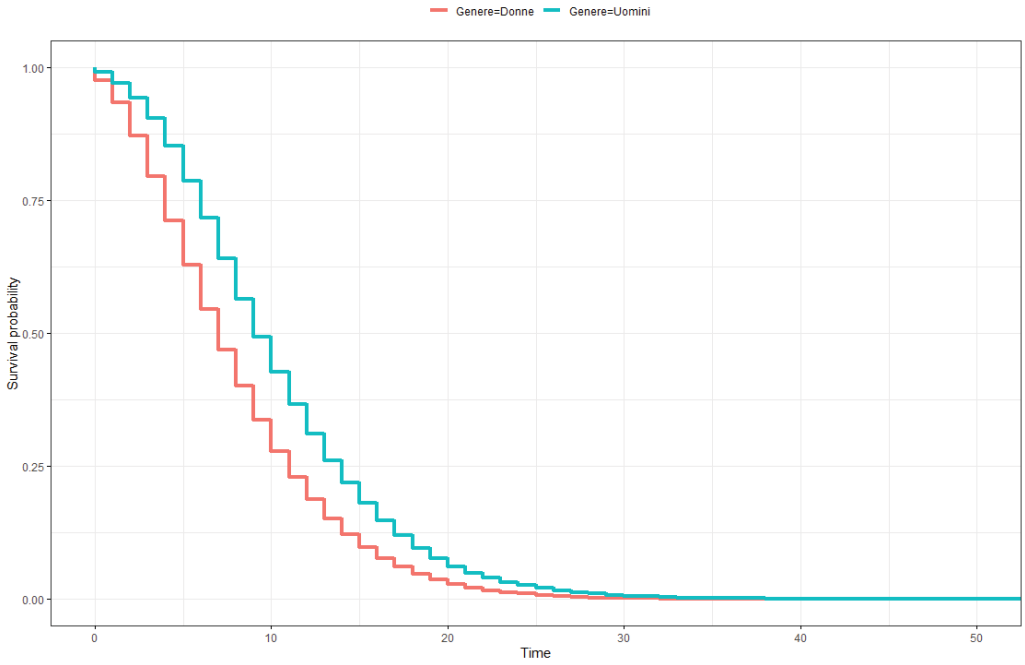


Figura 4.5: Curva di Kaplan-Meier per genere.

	N	**Observed**	**Expected**	$(O-E)^2/E$	$(O-E)^2/V$
Uomini	28800	28800	34471	933	2238
Donne	38201	38201	32530	989	2238
Chisq= 2238 on 1 degrees of freedom, p= <2e-16					

Tabella 4.12: *Log-rank test* - Confronto curve di sopravvivenza per genere.

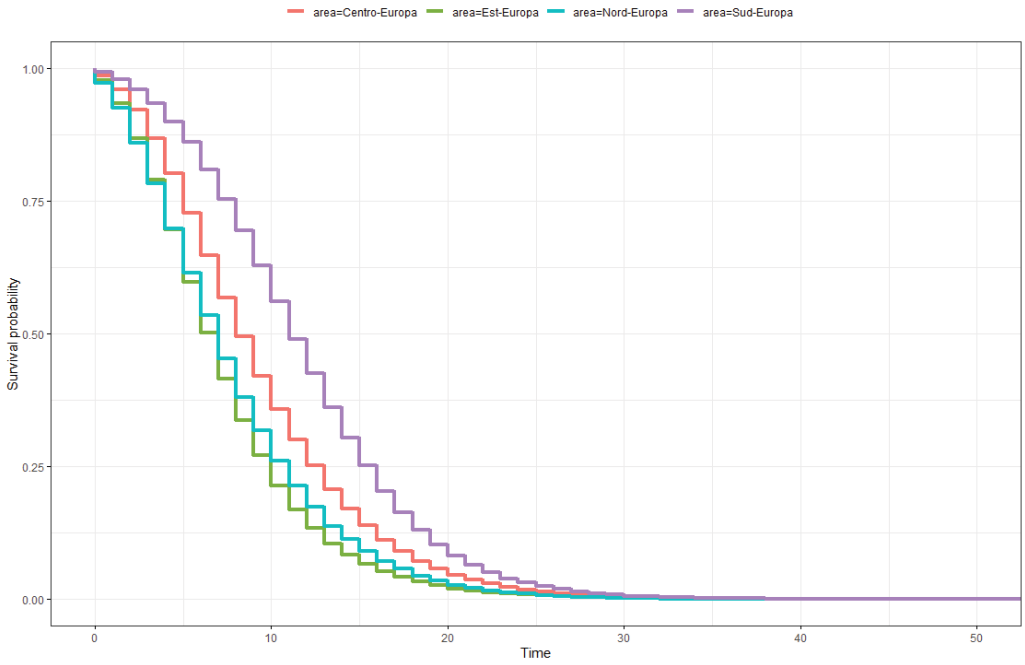


Figura 4.6: Curva di Kaplan-Meier per area geografica.

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Est-Europa	19210	19210	14336	1657.0	2437.4
Nord-Europa	13985	13985	11480	546.5	756.7
Centro-Europa	18632	18632	19364	27.7	44.8
Sud-Europa	15174	15174	21821	2024.6	3554.2
Chisq= 5042 on 3 degrees of freedom, p= <2e-16					

Tabella 4.13: *Log-rank test* - Confronto curve di sopravvivenza per area geografica.

Area geografica

Nella Figura 4.6 sono rappresentate le curve di sopravvivenza di Kaplan-Meier divise per area geografica.

Il rischio di anticipare la nascita del primo figlio diminuisce nelle aree geografiche Sud, Centro e Nord-Europa, rispetto ai Paesi dell'Est-Europa. In altre parole, i Paesi dell'Est-Europa sono più a rischio di avere una differenza, tra fine istruzione e nascita primo figlio, minore.

In Tabella 4.13 è riportato il *log-rank test*: anche in questo caso il p-value suggerisce che le curve di sopravvivenza per area geografica sono significativamente diverse.

4.4 Discussione

Dalle analisi e dai modelli stimati è possibile trarre alcune interessanti conclusioni.

In primo luogo, è stato studiato il ruolo dell'istruzione nel tempo che intercorre tra la fine del percorso scolastico e la nascita del primo figlio. Il modello ha evidenziato che un maggior numero di anni passati a scuola porta gli individui a ridurre, poi, il tempo che intercorre tra la fine dell'istruzione e la prima gravidanza. Inoltre, lo stesso modello ha evidenziato come lo stato lavorativo alla nascita del primo figlio contribuisca a spiegare il tempo che intercorre tra i due eventi. Per chi è casalingo/a o svolge altri lavori che non rientrano nelle categorie considerate dalla variabile, aumenta il rischio di ridurre il tempo tra i due eventi, rispetto ai disoccupati. Essere occupati indipendenti, invece, riduce il rischio di sperimentare l'evento rispetto ad essere disoccupati. Infine, la categoria *Occupato dipendente* non è risultata significativa.

Un ulteriore approfondimento, però, ha portato in luce come tale fenomeno sia legato all'aumento nel corso degli anni del livello di istruzione.

Infatti, è stato mostrato come la media delle persone che studiano più anni sia aumentata nel corso del tempo. In particolare, le persone intervistate in questo studio fanno parte di coorti di nascita dal 1925 fino al 1967.

Accanto a questo, si è indagato su un ulteriore aspetto, ossia l'età al primo figlio: è emerso che, nonostante nel corso degli anni il titolo di studio raggiunto dalle persone sia aumentato, non è cambiata in modo evidente, però, l'età alla prima gravidanza.

Quanto risultato porta a concludere che il titolo di studio non ha un effetto di accelerazione o decelerazione del tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio. Gli individui, che raggiungono livelli di istruzione sempre più elevati, e, di conseguenza, la loro età alla fine dell'istruzione aumenta, sono portati a velocizzare la nascita del primo figlio una volta terminato il percorso educativo, per motivi di capacità riproduttiva.

Da sottolineare che il campione di dati analizzato si riferisce a coorti di nascita che arrivano fino a metà degli anni '60: in questi anni ancora non è evidente l'effetto di un aumento dell'età al primo figlio e/o di una rinuncia ad averlo.

Successivamente, sono stati approfonditi due ulteriori variabili: il genere e l'area geografica.

Come ci si poteva aspettare, e proprio per motivi di fertilità, il rischio di sperimentare prima l'evento del primo figlio è maggiore per le donne, rispetto agli uomini. Per quanto riguarda l'area geografica, invece, il rischio di anticipare la nascita del primo figlio diminuisce nelle aree geografiche Sud, Centro e Nord-Europa, rispetto ai Paesi dell'Est-Europa. In altre parole, nei Paesi dell'Est-Europa, il tempo che intercorre dalla fine dell'istruzione e la nascita del primo figlio è minore rispetto agli altri Paesi Europei.

Questo studio presenta, tuttavia, alcuni limiti. Sono state considerate solo le persone che hanno avuto figli e non è stato, quindi, tenuto conto nelle analisi se l'istruzione abbia in qualche modo fatto aumentare le proporzione di persone che non ha mai avuto figli. Inoltre, non si tiene conto in maniera più approfondita del ruolo dei singoli Paesi, i quali vengono analizzati solo attraverso aree geografiche. Considerarli singolarmente potrebbe essere interessante perchè i vari Stati hanno sperimentato degli incrementi nel numero di anni di obbligo scolastico nel corso degli anni, e tali modifiche, però, sono avvenute in tempi e modi diversi tra i Paesi. Infine, l'assunzione di rischi proporzionali del modello di Cox non viene rispettata, per questo motivo tale soluzione non è stata ulteriormente utilizzata.

Conclusioni

Obiettivo di questo studio è l'analisi del tempo tra la fine dell'istruzione e la nascita del primo figlio in Europa. Quello che si è voluto indagare in questo elaborato non è tanto l'età alla prima maternità - argomento ampiamente discusso in letteratura-, ma la durata tra i due eventi appena menzionati: la fine dell'istruzione e la nascita del primo figlio. In particolare, si è voluto comprendere se c'è stata una accelerazione o una decelerazione di questo tempo in base agli anni di studio effettuati.

Le tecniche statistiche utilizzate in questo studio sono state i modelli per dati di durata, i quali hanno permesso di gestire le criticità del fenomeno e si sono rivelati adeguati a perseguire gli obiettivi iniziali.

Dall'analisi dei dati svolta si può concludere che, nel campione analizzato, che comprende gli anni dal 1925 fino al 1967, il livello di istruzione raggiunto è aumentato nel corso del tempo.

Accanto a questo, nonostante nel corso degli anni il titolo di studio raggiunto dalle persone sia aumentato, non è cambiata, in modo evidente, l'età alla prima gravidanza. Questo fenomeno porta a suggerire che ci sia stato un effetto di "adattamento" per avere il primo figlio, per motivi di capacità riproduttiva.

Inoltre, sempre per motivi di fertilità, il rischio di sperimentare l'evento del primo figlio è maggiore per le donne, rispetto agli uomini.

Per quanto riguarda l'area geografica, invece, il rischio di anticipare la nascita del primo figlio diminuisce nelle aree geografiche Sud, Centro e Nord-Europa, rispetto ai Paesi dell'Est-Europa. In altre parole, nei Paesi dell'Est-Europa, il tempo che intercorre dalla fine dell'istruzione alla nascita del primo figlio è minore rispetto agli altri Paesi Europei.

In conclusione, non si verifica un'accelerazione o una decelerazione del tempo che intercorre tra la fine dell'istruzione e la nascita del primo figlio in Europa a causa dell'istruzione: per il campione analizzato, che riguarda le coorti di nascita dal 1925 al 1967, il titolo di studio - o, più precisamente, gli anni di istruzione effettuati - non sono la mera causa di una anticipazione o meno del tempo che intercorre tra i due eventi. Gli individui, che raggiungono livelli di istruzione sempre più elevati, e, di conseguenza, la loro età alla fine dell'istruzione aumenta, sono portati a velocizzare la nascita del primo figlio una volta terminato il percorso educativo, per motivi di capacità riproduttiva.

Basandoci sui risultati di questo studio, si può suggerire, per eventuali politiche a supporto della fecondità in Europa, di diminuire il numero di anni di istruzione: ad esempio, diminuendo gli anni per la laurea o anticipando l'inizio della scuola dell'obbligo, in particolare nei Paesi del Nord, Centro e Sud-Europa. Lo studio ha considerato solo le persone che hanno avuto figli e non è stato, quindi, tenuto conto nelle analisi se l'istruzione abbia in qualche modo fatto aumentare le proporzione di persone che non ha mai avuto figli. Ciò nonostante, con la riduzione degli anni scolastici, si potrebbe verificare un'anticipazione nella nascita del primo figlio, evitando una potenziale rinuncia ad esso per la cosiddetta *fecondità involontaria* (Rowland, 2007): questo fenomeno avviene quando la gravidanza viene ritardata fino al punto in cui essa diviene improbabile o impossibile a causa di una ridotta fertilità individuale e/o di coppia, nel qual caso il rinvio volontario viene trasformato in infecondità involontaria.

Infine, utilizzando i dati SHARE, in continuo aggiornamento, si possono valutare ulteriori sviluppi del presente studio. In particolare, si potrebbe analizzare la differenza di anni tra fine istruzione e nascita del primo figlio nelle coorti di nascita più recenti, mettendo in luce eventuali similitudini e differenze con i risultati presentati in questo elaborato.

Appendice A

Codice R

A.1 Variabili ricostruite

```
# Variabile eta al primo figlio: eta_figlio
dati$eta_figlio<-NA
for (i in 1:81060) {
  for (j in 275:340) { #ychdstate columns
    if (is.na(dati$eta_figlio[i]) & !dati[i,j]=="0" & !dati[i,j]==".a")
      dati$eta_figlio[i] = substr(colnames(dati[j]), 10,11)
  }
}
dati$eta_figlio<-as.numeric(dati$eta_figlio)

# Variabile eta a fine istruzione: eta_fine_istr
dati$eta_fine_istr<-NA
#if workstate is 6 (6 = "full-time education")
for (i in 1:81060) {
  for (j in 11:76) { #workstate columns
    if(!is.na(dati[i,j]) & dati[i,j]=="7")
      dati$eta_fine_istr[i] = substr(colnames(dati[j]), 10,11)
  }
}
dati$eta_fine_istr<-as.numeric(dati$eta_fine_istr)
```

```
sum(is.na(dati$eta_fine_istr))

#if eta_fine_istr is NA: set with age at starting school for each country
dati$inizio_scuola<-6
for (i in 1:81060) {
  if(dati$country[i]==51 |dati$country[i]==35 | dati$country[i]==47|
     dati$country[i]==57| dati$country[i]==29 | dati$country[i]==48 |
     dati$country[i]==55 )
    dati$inizio_scuola[i] = 7
  if(dati$country[i]==59)
    dati$inizio_scuola[i] = 5
}
for (i in 1:81060) {
  if (is.na(dati$eta_fine_istr[i]))
  dati$eta_fine_istr[i] = dati$raedyrs[i] + dati$inizio_scuola[i]
}

# Variabile anni intercorsi tra fine istruzione e nascita primo figlio:
    diff
dati$diff<-NA
for (i in 1:81060) {
  dati$diff[i] = dati$eta_figlio[i] - dati$eta_fine_istr[i]
}

# Variabile stato lavorativo alla nascita del primo figlio: stato_lav
dati$stato_lav<-NA
for (i in 1:81060) {
  for (j in 77:142) { #hwrkstate columns
    if( !is.na(dati$eta_figlio[i]) & dati$eta_figlio[i] == substr(
        colnames(dati[j]), 10,11))
      dati$stato_lav[i] = dati[i, j]
  }
}
```

```
# Variabile stato partnership alla nascita del primo figlio: stato_
    partnership
dati$stato_partnership<-NA
for (i in 1:81060) {
  for (j in 143:208) { #prtnstate columns
    if(!is.na(dati$eta_figlio[i]) & dati$eta_figlio[i] == substr(colnames
        (dati[j]), 10,11))
      dati$stato_partnership[i] = dati[i, j]
  }
}

# Variabile stato abitativo alla nascita del primo figlio: stato_abit
dati$stato_abit<-NA
for (i in 1:81060) {
  for (j in 406:471) { #rsidstate columns
    if(!is.na(dati$eta_figlio[i]) & dati$eta_figlio[i] == substr(colnames
        (dati[j]), 10,11))
      dati$stato_abit[i] = dati[i, j]
  }
}
```

A.2 Grafici

```
#Anni intercorsi tra istruzione e nascita primo figlio - Confronto uomini
    e donne
ggplot(df, aes(x=raedyrs6, y=diff, fill=ragender)) +
  geom_boxplot()+
  scale_fill_manual(values = c("lightblue", "pink"),labels = c("Uomo", "
      Donna"))+
  labs(fill = "Sesso")+
  ggtitle("")+
```

```
  labs(x="Anni di istruzione", y="Anni intercorsi da fine istruzione a
      primo figlio")+
  theme(axis.text.x= element_text(size=11))+
  theme(axis.text.y= element_text(size=11))+
  theme(axis.title.x = element_text(size = rel(1.2), angle = 00))+
  theme(axis.title.y = element_text(size = rel(1.2), angle = 90))

#Anni intercorsi tra istruzione e nascita primo figlio - Confronto tra
    paesi
df %>%
  ggplot(aes(x = reorder(country, diff, fun=mean, na.rm=T) , y = diff ))
      +
  geom_boxplot(aes(fill=reorder(country, diff, fun=mean, na.rm=T))) +
  labs(x = '', y = 'Anni tra istruzione e primo figlio') +
  coord_flip() +
  ggtitle("")+
  theme_bw()+
  theme(legend.position = "none")+
  theme(axis.text.x= element_text(size=11))+
  theme(axis.text.y= element_text(size=11))+
  theme(axis.title.x = element_text(size = rel(1.2), angle = 00))+
  theme(axis.title.y = element_text(size = rel(1.2), angle = 90))

#Eta al primo figlio per coorte di nascita
cdat <- ddply(df, "coorti", summarise, rating.mean=median(eta_figlio))
ggplot(df, aes(x = eta\_figlio, fill=coorti)) +
  geom_histogram(binwidth=1, position = "identity", colour= "black") +
  labs(x = "Eta al primo figlio", y="")+
  theme(legend.title = element_blank())+
  facet_grid(coorti ~ .)+
  geom_vline(data=cdat, aes(xintercept=rating.mean,  colour=coorti),
             linetype="dashed", size=1)

#Distribuzione degli anni di istruzione per ogni coorte di nascita.
```

```
ggplot(df, aes(x = raedyrs, fill=coorti)) +
  geom_histogram(binwidth=1, position = "identity", colour= "black") +
  facet_grid(coorti ~ .) +
  theme(legend.title = element_blank())+
  labs(x = "Anni di istruzione", y="")+
  geom_vline(data=cdat, aes(xintercept=rating.mean,  colour=coorti),
             linetype="dashed", size=1)
```

A.2.1 Analisi delle sequenze

```
# Creo sequenze lavoro
#co<-hcl.colors(8, palette='RdYlBu')
Hwork.labels <- c("Eta non attribuibile" ,"Occupato dipendente",  "
    Occupato indipendente",  "Disoccupato",  "Casalingo/a",  "Pensionato"
    ,  "Studente",  "Altro")
Hwork.scode <- c("AN", "EM",  "SE",  "UN",  "FA",  "RE",  "ED",  "OT")
Hwork.seq.zero <- seqdef(zero_figli, 77:142, states = Hwork.scode, labels
     = Hwork.labels)
Hwork.seq.uno <- seqdef(unopiu_figli, 77:142, states = Hwork.scode,
    labels = Hwork.labels)

# Grafico
par(mfrow=c(2,2))
seqdplot(Hwork.seq.zero, with.legend = F, border = NA,main = "Nessun
    figlio")
seqdplot(Hwork.seq.uno, with.legend = F, border = NA, main = "Uno o piu
    figli")
seqlegend(Hwork.seq.zero, cex = 1.1)

# Sequenze partnership
prtn.labels <- c("Eta non attribuibile" , "Vive solo", "Vive con il
    partner")
prtn.scode <- c("AN","AL", "PA")
```

```
prtn.seq.zero <- seqdef(zero_figli, 143:208, states = prtn.scode, labels
    = prtn.labels)
prtn.seq.uno <- seqdef(unopiu_figli, 143:208, states = prtn.scode, labels
     = prtn.labels)

#Grafico
par(mfrow=c(2,2))
seqdplot(prtn.seq.zero, with.legend = F, border = NA, main = "Nessun
    figlio")
seqdplot(prtn.seq.uno, with.legend = F, border = NA, main = "Uno o piu
    figli")
seqlegend(prtn.seq.zero, cex = 1.3)
```

A.2.2 Curve di Kaplan-Meier

```
# K-M diviso per anni di istruzione
k1=survfit(Surv(df$diff)~df$anni\_istr)
ggsurvplot( k1,  data = df,  size = 1.5,  legend.title="",  ggtheme =
    theme_bw())

# K-M diviso per coorte
k2=survfit(Surv(df$diff)~df$coorte)
ggsurvplot( k2,  data = df,  size = 1.5,  legend.title="",  ggtheme =
    theme_bw())

# K-M diviso per genere
k3=survfit(Surv(df$diff)~df$Genere)
ggsurvplot( k3,  data = df,  size = 1.5,  legend.title="",  ggtheme =
    theme_bw())

# K-M diviso per area geografica
k4=survfit(Surv(df$diff)~df$area)
ggsurvplot( k4,  data = df,  size = 1.5,  legend.title="",  ggtheme =
    theme_bw())
```

```
#Log-rank test
survdiff(Surv(df$diff)~df$anni\_istr)
survdiff(Surv(df$diff)~df$coorte)
survdiff(Surv(df$diff)~df$Genere)
survdiff(Surv(df$diff)~df$area)
```

A.3 Modelli stimati

```
#Modello baseline
mod<-coxph(Surv(diff) ~ raedyrs6+coorti+ragender+area+stato_lav, data =
    df)
summary(mod)

#Modello cox univariato - anni istruzione
m2 <- coxph(Surv(diff) ~ raedyrs6, data = df)
summary(m2)

#Un modello per ogni coorte di nascita.
c1<-df[df$coorti=="1925—1935", ]
c2<-df[df$coorti=="1936—1940", ]
c3<-df[df$coorti=="1941—1945", ]
c4<-df[df$coorti=="1946—1950", ]
c5<-df[df$coorti=="1951—1955", ]
c6<-df[df$coorti=="1956—1960", ]
c7<-df[df$coorti=="1961—1967", ]

mc1<-coxph(Surv(diff) ~ raedyrs6, data = c1)
mc2<-coxph(Surv(diff) ~ raedyrs6, data = c2)
mc3<-coxph(Surv(diff) ~ raedyrs6, data = c3)
mc4<-coxph(Surv(diff) ~ raedyrs6, data = c4)
mc5<-coxph(Surv(diff) ~ raedyrs6, data = c5)
```

```
mc6<-coxph(Surv(diff) ~ raedyrs6, data = c6)
mc7<-coxph(Surv(diff) ~ raedyrs6, data = c7)


#Diagnostica del modello di Cox
cox.zph(mod)
```

Bibliografia

B. Arpino, G. Esping-Andersen, and L. Pessin. How do changes in gender role attitudes towards female employment influence fertility? A macro-level analysis. *European Sociological Review*, 31(3):370–382, 2015.

E. Beaujouan, Z. Brzozowska, and K. Zeman. Childlessness trends in twentieth-century europe: Limited link to growing educational attainment. Technical report, Vienna Institute of Demography Working Papers, 2015.

G. S. Becker. A theory of marriage: Part II. *Journal of political Economy*, 82(2, Part 2):S11–S26, 1974.

M. Bergmann, A. Scherpenzeel, and A. Börsch-Supan. Share wave 7 methodology: Panel innovations and life histories. *Munich: Munich Center for the Economics of Aging*, 2019.

A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *International journal of epidemiology*, 42(4):992–1001, 2013.

J. C. Caldwell. Mass education as a determinant of the timing of fertility decline. *Population and development review*, pages 225–255, 1980.

S. Cantalini. Formazione della famiglia, fecondità e stratificazione sociale nell'Italia dal dopoguerra a oggi. Master's thesis, Università degli Studi di Milano, 2017.

European Commission. Demography report-short analytical web note, 2015.

European Commission. Compulsory education in europe - 2018/19, 2018a.

European Commission. Bye bye parents: when do young europeans flee the nest?, 2018b.

L. Gangadharan and P. Maitra. The effect of education on the timing of marriage and first birth in pakistan. *Journal of Quantitative Economics*, 1(1):114–133, 2003.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282): 457–481, 1958.

J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

E. Mattiollo. Analisi robusta della sopravvivenza nello studio del mesotelioma maligno. Master's thesis, Università degli Studi di Padova, 2012.

A. R. Miller. The effects of motherhood timing on career path. *Journal of population economics*, 24(3):1071–1100, 2011.

M. Ni Bhrolchain and E. Beaujouan. Fertility postponement is largely due to rising educational enrolment. *Population studies*, 66(3):311–327, 2012.

C. Nicoletti and M. L. Tanturri. Differences in delaying motherhood across European countries: Empirical evidence from the ECHP. *European Journal of Population/Revue Européenne de Démographie*, 24(2):157–183, 2008.

L. Pace and A. Salvan. *Introduzione alla statistica: Inferenza, verosimiglianza, modelli.-2001.-xvi, 422 p.* Cedam, 1996.

E. T. Pascarella and P. T. Terenzini. *How College Affects Students: A Third Decade of Research. Volume 2.* ERIC, 2005.

L. Persson. Trend reversal in childlessness in sweden. *Work session on demographic projections*, pages 129–135, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

D. T. Rowland. Historical trends in childlessness. *Journal of family Issues*, 28(10):1311–1337, 2007.

T. Sobotka. Pathways to low fertility: European perspectives. *Expert Paper*, 8, 2013.

M. L. Tanturri. "Una rivoluzione tutta per sé": le donne e il cambiamento demografico. *Polis*, 23(2):309–320, 2009.

UNESCO Institute for Statistics. *International standard classification of education: ISCED 2011*. UNESCO Institute for Statistics Montreal, 2012.

D. J. Van de Kaa. Europe's second demographic transition. *Population bulletin*, 42(1):1–59, 1987.

M. L. Zanier. Il declino della fecondità nei paesi occidentali. *Polis*, 16(3): 347–374, 2002.