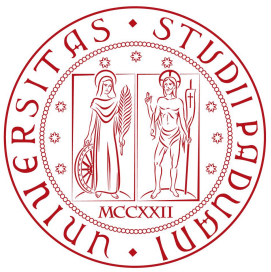
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche



Ritorno della fertilità dopo il parto: analisi dei fattori fisiologici e metabolici

Relatore: dott. Bruno Scarpa
Dipartimento di Scienze Statistiche

Laureando: Fabio Barbaro
Matricola n. 1106295

Anno Accademico 2016/2017

Indice

Capitolo 1

Introduzione

Nel periodo immediatamente successivo al parto, si osserva una naturale interruzione della normale fertilità femminile. Questo periodo di amenorrea post-parto è di lunghezza piuttosto variabile: la comprensione di quali fattori possano influenzare la sua durata è fondamentale al fine di permettere alla donna di spaziare tra le gravidanze ed occuparsi al meglio dello sviluppo del bambino. È noto come l'allattamento esclusivo al seno sia un fattore preponderante nel ritardare il ritorno della fertilità dopo il parto, risulta inoltre essere il modo più sano per alimentare il bambino e garantirgli un sano sviluppo.

Alcuni ricercatori della *Marquette University* nel Michigan hanno formulato un protocollo per aiutare le donne il più possibile nel rilevare quando stia per avvenire il ritorno della fertilità: questo si basa sulle rilevazioni elettroniche di due ormoni tipicamente associati al ritorno alla fertilità, l'ormone Estrone-3-Glucuronide ($E3G$) e l'ormone luteinizzante (LH). Il dispositivo di monitoraggio elettronico, utilizzato giornalmente, può segnalare tre tipi di output diversi: un picco di ormone luteinizzante, un alto valore di E3G, oppure non segnalare attività ormonale rilevante.

Nel nostro studio siamo interessati a comprendere quali fattori influenzino la durata del periodo di amenorrea post-parto, ponendo particolare attenzione sull'effetto dell'allattamento e dei due biomarcatori ormonali descritti in precedenza. È inoltre di grande interesse comprendere quali siano i fattori che influiscono sul fatto di non produrre abbastanza ormone LH durante un'ovulazione da permettere al rilevamento elettronico di segnalare un picco di questo biomarcatore, dato che il protocollo di pianificazione familiare si basa soprattutto sulla rilevazione di un'elevata presenza di ormone luteinizzante per segnalare il ritorno alla fertilità.

Se riuscissimo a comprendere al meglio il meccanismo di ritorno della

fertilità rispondendo ai due quesiti di interesse potremmo rendere il protocollo ancora più efficace: sarebbe quindi possibile pianificare le gravidanze successive con maggiore efficacia.

La peculiarità di questo studio è di essere il primo in cui viene valutato l'effetto dell'allattamento esclusivo al seno sulla durata della fase di amenorrea post-parto al netto dei due fattori ormonali sopracitati.

Nel Capitolo 2 verrà spiegato il problema più nel dettaglio, sottolineando il rapporto tra allattamento e ritorno della fertilità. Verrà spiegata la modalità con cui le donne sono entrate a far parte dello studio, come sono stati raccolti i dati e gli obiettivi finali dello studio. Nel capitolo 3 verranno esaminate nel dettaglio le variabili utilizzate nell'analisi: le durate di interesse, il meccanismo di censura, i dati mancanti presenti nel dataset e le variabili esplicative utilizzate, spiegando nel dettaglio le operazioni eseguite su di esse prima di inserirle nei modelli statistici. Nel Capitolo 4 verranno effettuate le analisi statistiche per rispondere alla prima delle due domande di interesse dello studio, ossia su quali siano i fattori che influenzano il ritorno della fertilità post-parto. Verranno discussi i modelli utilizzati ed i metodi per trattare i valori mancanti. Nel Capitolo 5 verranno esposti i modelli statistici utilizzati per rispondere alla seconda delle domande di interesse, ossia su quali siano i fattori che influenzano il fatto di osservare almeno un picco di ormone luteinizzante durante la fase di amenorrea post-parto.

I risultati ottenuti dalle analisi sono perfettamente in linea con le conoscenze scientifiche pregresse ottenute da altri studi precedenti effettuati nello steso ambito. Abbiamo ottenuto una conferma del ruolo preponderante dell'allattamento esclusivo al seno sul ritorno della fertilità, abbiamo notato inoltre un forte effetto dei fattori ormonali sul rischio di sperimentare questo evento ed una certa interazione tra queste due quantità. Abbiamo osservato come al crescere dell'indice di massa corporea aumenti la probabilità di osservare picchi di ormone LH, inoltre, il fatto di osservare un'elevata quantità di ormone E3G è associato ad un aumento della produzione di ormone luteinizzante.

Capitolo 2

Descrizione dell'analisi

2.1 Descrizione dell'analisi

Il problema affrontato in questa tesi di laurea ci è stato fornito dalla facoltà di Infermieristica della *Marquette University* a Milwaukee, nel Michigan. Insieme alla spiegazione del problema ed all'esposizione degli scopi dello studio, ci è stato fornito il *dataset* su cui effettuare le analisi, insieme a del materiale informativo e qualche riferimento bibliografico da consultare.

È ben noto come il normale svolgimento del ciclo mestruale venga soppresso immediatamente dopo il parto per un certo periodo di tempo (la cosiddetta fase di "amenorrea post parto"). La durata di questa fase di infertilità è molto variabile a seconda dei casi (Bouchard *et al.*, 2013), ed è di grande importanza comprendere al meglio i meccanismi che determinano la durata di questo periodo biologico.

Il latte materno si è rivelato essere il modo più sano di alimentare un neonato (Gartner *et al.*, 2005): sono noti gli effetti positivi che questo regime alimentare ha sulla protezione del bambino dalle infezioni virali, sul corretto sviluppo delle sue capacità cognitive, sulla prevenzione del diabete e di altre malattie legate ad una scorretta alimentazione (Victora *et al.*, 2016). Numerosi studi precedenti dimostrato inoltre come l'allattamento esclusivo al seno sia un fattore determinante nel ritardare il ritorno alla fertilità, almeno nel periodo compreso tra i sei ed i nove mesi successivi al parto (McNeilly, 2002).

La necessità di avere il controllo sui tempi di intercorrenza tra una gravidanza e l'altra per una neo madre risulta fondamentale ai fini della qualità dello sviluppo del bambino. L'utilizzo di contraccettivi orali per spaziare le gravidanze è indubbiamente molto efficace, tuttavia in alcuni casi l'assunzione di questi farmaci potrebbe essere dannoso per

la qualità del latte materno e portare ad una precoce interruzione della sua produzione (McCann *et al.*, 1981). Se comprendessimo al meglio i meccanismi di ritorno alla fertilità si potrebbero formulare protocolli che permettano di alimentare il bambino nel modo più sano, ed allo stesso tempo di evitare gravidanze indesiderate.

L'interesse principale dei ricercatori che ci hanno fornito il problema è quindi di comprendere quali variabili abbiano influenza sul ritorno della fertilità dopo il parto, ponendo particolare enfasi sulle modalità di allattamento ed alimentazione del bambino.

La peculiarità di questa analisi risiede nel fatto di valutare l'effetto che tipologie di variabili molto diverse tra loro hanno sul tempo di ritorno alla fertilità: stando alle informazioni forniteci, nessuno studio precedente aveva mai cercato di comprendere il meccanismo del ritorno della fertilità combinando gli effetti del tipo di alimentazione del bambino coi fattori fisiologici e metabolici della madre, inclusa l'informazione relativa a biomarcatori ormonali che verranno spiegati più nel dettaglio nella sezione 2.2.

Il campione analizzato è composto da 93 donne americane di età compresa tra i 23 ed i 43 anni che abbiano partorito tra le sei e le otto settimane precedenti all'ingresso nello studio. Ciascuna di esse ha preso parte ad un particolare protocollo di pianificazione familiare naturale, denominato dai ricercatori dell'università del Michigan *"Marquette Method"* (Fehring *et al.*, 2008), basato su un metodo di monitoraggio ormonale che verrà presentato più nel dettaglio nella sezione successiva.

Per ogni donna sono state raccolte informazioni riguardanti la durata della fase di amenorrea post parto e dei cicli mestruali immediatamente successivi al ritorno della fertilità. Il numero di cicli monitorati varia da soggetto a soggetto, ed oscilla tra un minimo di nessun ciclo registrato ad un massimo di sei cicli monitorati. Il periodo del ritorno alla fertilità è stato denominato "Ciclo 0", mentre i periodi successivi rispettivamente "Ciclo 1", "Ciclo 2", "Ciclo 3", "Ciclo 4" ,"Ciclo 5", "Ciclo 6". Nelle nostre analisi, i cicli successivi al ritorno della fertilità non sono di interesse, ma ci sono stati comunque forniti in quanto già presenti nel dataset originale.

2.2 Il monitoraggio elettronico del *"Marquette Method"*

A ciascuna donna è stata fornita una particolare apparecchiatura elettronica, denominata "EHFM" (dall'inglese *Electronic Home Fertility Monitor*), in grado di rilevare la presenza di due particolari tipi di ormoni presenti nelle urine: l'ormone Estrone-3-Glucuronide (*E3G*), e l'ormone Luteinizzante (*LH*) (Fehring *et al.*, 2007). La produzione di E3G è regolata dalla presenza di estrogeni ed aumenta al crescere del loro numero; questo fenomeno avviene tipicamente nel periodo precedente all'ovulazione. L'ormone LH, invece, è prodotto dall'ipofisi: una modesta quantità di questo ormone è sempre presente nel collo uterino, tuttavia si registra un notevole incremento della sua produzione alcuni giorni prima dell'ovulazione. Questi due ormoni risultano essere particolarmente utili come indicatori del fatto che l'ovulazione stia per avvenire (WorldHealthOrganization, 1981): questo concetto è alla base di svariati metodi per l'individuazione del miglior giorno per il concepimento, come ad esempio in Royston (1991), in Behre *et al.* (2000) o in Robinson *et al.* (2007).

Il protocollo prevede l'analisi giornaliera delle prime urine mattutine; il responso del dispositivo può essere di tre tipi, ed è codificato nelle tre classi seguenti, che indicano il livello di fertilità nella giornata corrente:

- LOW: non è stata rilevata la presenza significativa di nessuno dei due ormoni sopracitati.
- HIGH: è stato rilevato un notevole aumento dell'ormone E3G. Questo avviene nel caso in cui la misura della quantità di questo ormone superi una certa soglia di tolleranza.
- PEAK: è stato rilevato un notevole aumento di ormone LH. In analogia col caso precedente, l'EHFM fornisce questo esito nel caso in cui la soglia di tolleranza relativa alla quantità di ormone luteinizzante osservata venga oltrepassata.

Per semplicità, nel seguito di questa tesi chiamerò i valori *LOW*, *HIGH* e *PEAK*, semplicemente "basso/i", "alto/i" e "picco/picchi".

Le donne appartenenti al campione considerato sono tutte provviste di *Smartphone*; i dati relativi al monitoraggio ormonale sono stati inviati ai ricercatori mediante l'inserimento giornaliero in una particolare applicazione *online* del valore fornito dall'apparecchiatura EHFM, insieme ad un valore di intensità di sanguinamento nel caso in cui siano avvenute

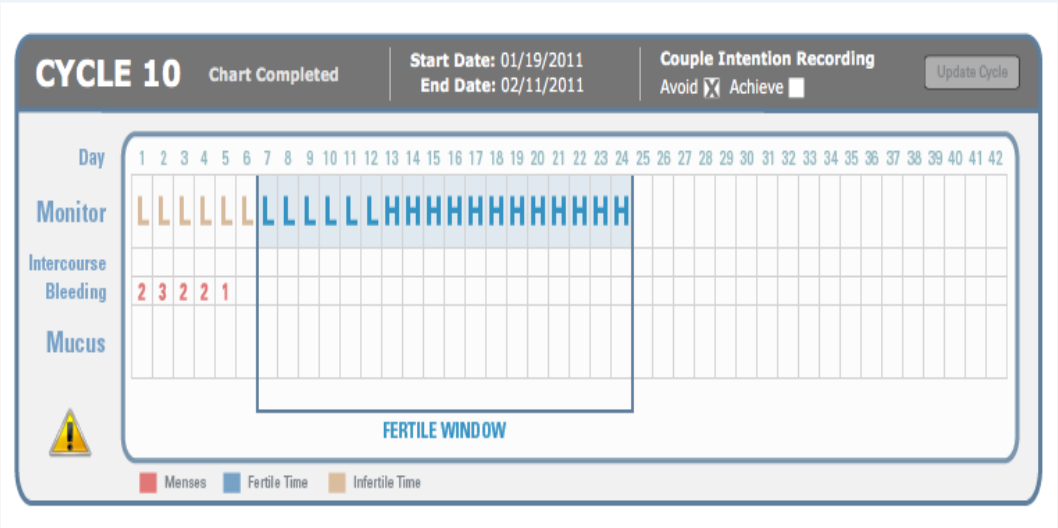


Figura 2.1: Interfaccia di inserimento dei dati ottenuti dall'EHFM.

delle perdite di sangue durante la giornata. Nella figura 2.1 è riportato un esempio di come funziona questa applicazione: nella sezione "Monitor" viene inserito l'esito dell'analisi EHFM; se ci sono state perdite di sangue, invece, ne viene annotata l'intensità nella sezione "*Intercourse Bleeding*", con codifica 1 per intensità leggera, 2 per intensità media e 3 per elevata intensità.

Questi ultimi valori verranno utilizzati nel determinare l'effettivo arrivo della prima mestruazione fertile e del conseguente ritorno alla fertilità: il Ciclo 0 si può definire concluso nel caso in cui i valori di intensità di sanguinamento abbiano descritto un *pattern* crescente-decrescente di lunghezza non inferiore ai due giorni, la cui somma sia almeno pari a 5. Secondo quanto comunicatoci dai ricercatori della Marquette University, è altamente improbabile osservare il ritorno alla fertilità in un periodo inferiore ai 56 giorni dopo il parto.

L'individuazione di un picco segnalato dal dispositivo risulta particolarmente importante, poiché la stima del giorno di ovulazione, denominata "EDO" (dall'inglese *Estimated Day of Ovulation*), è definita come l'ultimo valore di picco documentato dal dispositivo prima del ritorno della mestruazione successiva. Tipicamente, nel Ciclo 0 i giorni in cui si registrano il primo alto ed il primo picco precedono quello del ritorno della fertilità, tuttavia si è notato come, in alcuni soggetti, la data delle prime mestruazioni preceda quella del primo picco osservato dall'EHFM. I ricercatori sono interessati a comprendere se e quali variabili possano influenzare il fatto di aver già sperimentato il ritorno alla fertilità senza aver prima osservato almeno un valore di picco nel periodo precedente.

2.3 Descrizione del dataset

Nel dataset fornitoci sono state registrate diverse variabili relative alle donne partecipanti allo studio, raggruppabili principalmente in cinque categorie diverse:

- Variabili relative alle durate: per ogni donna sono state registrate la data del parto e la data della prima mestruazione fertile, che contrassegna la fine del Ciclo 0 ed il ritorno della fertilità. Ci è stata anche fornita la variabile che registra la differenza in giorni tra queste due date.

- Variabili relative alla sfera sociale: nel dataset sono documentati lo stato civile della donna, con relativa durata espressa in anni dell'eventuale matrimonio, la religione, l'etnia di appartenenza e gli anni di istruzione.

- Variabili fisiologiche: sono riportate l'età, l'altezza, il peso e l'indice di massa corporea ("BMI") relativi ad ogni donna, registrati al momento dell'ingresso nello studio. Sono anche forniti il numero di gravidanze che la donna ha sperimentato, compresa l'attuale, ed il numero di figli attualmente in vita. L'altezza delle donne è espressa in piedi ("*feets*"), mentre il peso sotto forma di libbre ("*pounds*").

- Fattori metabolici: due variabili relative alle modalità di alimentazione del neoneato sono di interesse: il fatto che il figlio sia alimentato esclusivamente con allattamento al seno (espressa mediante una variabile dicotomica "Allattamento Totale" o "Allattamento Parziale") ed il mese d'età del bambino in cui sono stati introdotti nella sua alimentazione liquidi od alimenti diversi dall'esclusivo latte materno, la variabile "Supplementi".

- Fattori relativi ai biomarcatori ormonali: queste variabili provengono unicamente dall'*output* giornaliero dall'EHFM. Nel dataset sono registrate le date relative al primo valore *HIGH*, al primo valore *PEAK* ed al giorno di ovulazione stimato rilevati durante la fase di amenorrea post-parto per i diversi soggetti. Sono registrati inoltre il numero di bassi, di alti e di picchi registrati dall'apparecchio durante il Ciclo 0. Abbiamo inoltre a nostra disposizione una variabile dicotomica che dice se prima del ritorno della fertilità sia stata stimata un'ovulazione, ossia se tra i valori forniti dall'EHFM sia stato registrato almeno un valore di picco. Sono anche presenti

due variabili che ci dicono se le mestruazioni che hanno marcato il ritorno alla fertilità siano state effettivamente fertili o meno, col relativo valore osservato della perdita di sangue.

- Variabili relative al monitoraggio: nel dataset sono specificati il giorno di inizio ed il giorno di fine del monitoraggio elettronico, con relativo numero di giorni monitorati.

Le caratteristiche delle singole variabili verranno trattate in maniera molto più approfondita nella sezione 3.5.

2.4 Obiettivi dell'analisi

Le domande d'interesse sono state già accennate in precedenza; in questa breve sezione, verranno esposte con maggiore chiarezza.

- La domanda principale dello studio è quella di valutare l'effetto che hanno i diversi fattori esplicativi sui tempi di ritorno alla fertilità. Particolare enfasi è posta sul valutare l'effetto dell'allattamento esclusivo al seno al netto delle altre variabili, soprattutto di quelle relative ai biomarcatori ormonali.

- La seconda domanda d'interesse dello studio è di valutare l'effetto dei diversi fattori sul fatto di causare durante il periodo di amenorrea post parto almeno un picco di ormone luteinizzante abbastanza intenso da essere rilevato dal dispositivo di monitoraggio elettronico EHFM.

Nelle analisi statistiche questi due quesiti d'interesse verranno trattati separatamente, formulando modelli statistici differenti per ogni domanda.

Capitolo 3

Pre-analisi: osservazioni, operazioni preliminari e statistiche descrittive

3.1 Operazioni preliminari sul dataset - pulizia dei dati

Prima di cominciare le analisi statistiche vere e proprie, sono state effettuate alcune operazioni preliminari di pulizia dei dati sul dataset originale. Innanzitutto, delle 97 osservazioni che originariamente componevano il dataset, una è stata esclusa in quanto presente due volte (anche se con qualche piccola differenza tra i valori di alcune variabili - evidente errore di trascrizione). Tre donne, invece, sono state eliminate dallo studio poiché non avevano altre informazioni da apportare all'analisi oltre alla data di nascita del bambino ed il mese in cui sono stati introdotti i "Supplementi".

In generale, nel dataset è stato individuato qualche sporadico errore di trascrizione: questi sono sempre stati corretti, dopo aver consultato i ricercatori che ci hanno fornito il problema. La variabile relativa alla durata della fase di amenorrea post-parto ci è stata fornita sia sotto forma di data che di durata espressa in giorni; stesso discorso per il primo alto, il primo picco e l'EDO relativi al Ciclo 0. Il fatto di avere a disposizione due formulazioni diverse che codificano la stessa variabile ci ha permesso di andare a correggere manualmente più di qualche errore di trascrizione osservato nel dataset, utilizzando molta cautela ed un po' di buon senso.

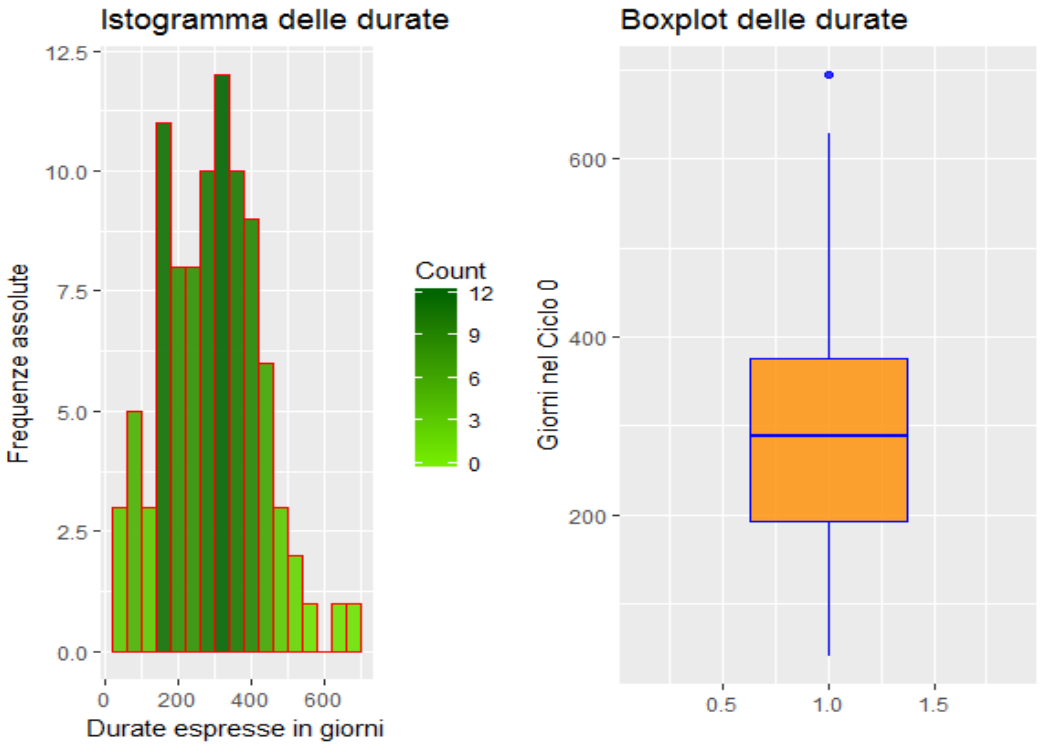


Figura 3.1: Istogramma e boxplot relativi alla distribuzione della variabile risposta.

La variabile indicatrice relativa al fatto di avere un EDO durante il Ciclo 0 presentava abbastanza discrepanze con le date forniteci dalle altre variabili presenti nel dataset: abbiamo corretto manualmente i valori nei casi in cui risultasse appropriato farlo.

3.2 La durata del periodo di amenorrea post-parto

Nella figura 3.1 sono riportati l'istogramma ed il *boxplot* relativi alla distribuzione della variabile relativa ai tempi di durata della fase di amenorrea post-parto. La media delle durate di questa variabile è approssimativamente pari a 291 giorni, con uno scarto quadratico medio pari a circa 135 giorni. Il valore minimo di giorni compresi tra la nascita del bambino è pari a 42, mentre il valore massimo registrato è pari a 694.

La variabile risposta ci è stata fornita nel dataset originale mediante due tipi di codifica diversa: sono riportate sia le due date di inizio e di fine del Ciclo 0 che la durata espressa direttamente in giorni. In gene-

rale si è preferito ottenere il tempo di durata mediante l'utilizzo della differenza tra le due date piuttosto che prendere direttamente il valore della durata espresso in maniera esplicita: ci è stato infatti comunicato dai ricercatori che la codifica in date risulta più affidabile rispetto alle durate in giorni. Quest'ultima variabile è stata inserita nel dataset facendo manualmente la differenza in giorni tra le due date, e in più di qualche caso è notata stata qualche discrepanza. Ci è stato confermato che si è trattato di errori di trascrizione.

3.3 Censure

Come nella grande maggioranza delle analisi statistiche relative a dati di durata, in questo studio certe durate associate ad alcune unità statistiche non terminano con l'evento di interesse, ma, semplicemente, ad un certo punto i soggetti escono dallo studio ed i ricercatori non sono più in grado di ottenere informazioni ulteriori. La presenza di dati censurati è uno dei problemi più tipici nell'ambito dei dati di sopravvivenza, ed è stato ampiamente trattato in letteratura (Klein e Moeschberger, 2005). Esistono diverse possibili tipologie di censura nell'ambito dei dati di durata: il tipo di censura descritto sopra è il più comune, ed è denominato "censura a destra". Tipicamente, la censura viene inserita nei modelli di sopravvivenza sotto forma di variabile dicotomica: all'i-esima unità statistica viene associato un indicatore C_i che assume valore 1 nel caso in cui il caso sia completo, ossia che l'uscita dall'insieme dei soggetti a rischio per quel particolare evento combaci effettivamente con l'evento finale di interesse, e valore 0 nell'eventualità in cui si sia verificata una censura.

Nel nostro studio le censure non ci sono state fornite esplicitamente sotto forma di variabili dicotomiche: è stato necessario infatti ricavarle a partire da altre variabili. Una sola unità tra i soggetti appartenenti allo studio risulta censurata. Non avendo a disposizione né la data della prima mestruazione fertile né la durata in giorni del Ciclo 0, abbiamo posto la variabile risposta relativa a questa unità come pari alla differenza tra l'ultima data relativa a qualche evento registrato dal monitoraggio ormonale e la data del parto: noi non abbiamo informazione su cosa sia successo dopo questa data, ma sappiamo che almeno fino a quel momento il soggetto faceva ancora parte dell'insieme delle unità a rischio. In particolare, l'ultimo evento registrato per questa donna è stato il ri-

levamento del primo alto ormonale, avvenuto nel sessantesimo giorno dopo la nascita del bambino.

3.4 Valori mancanti

Numerosi valori mancanti sono presenti nel dataset originale: tramite la stessa sigla *NA* sono indicate due tipologie di dati mancanti concettualmente differenti:

- Alcuni di questi sono mancanti perché effettivamente non avvenuti, come ad esempio le date dei valori di alto, picco, ed EDO nel caso in cui non ci siano stati alti o picchi ormonali durante il periodo di amenorrea post-parto.

- La maggior parte dei valori non osservati presenti nel dataset derivano dal fatto che i ricercatori non siano effettivamente riusciti ad ottenere le informazioni relative ai valori di alcune variabili per l'unità statistica in questione. Considerando solo quest'ultima categoria, circa il 15% dei valori del dataset risulta mancante, con maggior concentrazione di assenza di informazione nelle variabili relative all'output dell'EHFM.

Per alcuni soggetti è stato possibile ricavare dei valori mancanti tramite l'utilizzo di altre variabili complete: i valori mancanti relativi alla variabile "BMI" sono stati ricavati dalla relazione deterministica definita a partire dalle altre due, ossia $peso/altezza^2$, nel caso in cui le variabili "altezza" e "peso" per la stessa unità statistica fossero complete. Come già spiegato nella sezione 3.2, la variabile relativa alla durata del Ciclo 0 è espressa sia mediante le date di inizio e fine della fase di amenorrea che come durata in giorni, scegliendo di utilizzare la differenza tra due date come fonte di informazione. Tuttavia, nel caso in cui almeno una delle due date fosse mancante, la durata è stata stata recuperata dalla seconda tipologia di codifica, nel caso in cui non fosse mancante anche questo valore.

Nel paragrafo successivo, la sezione 3.5, le diverse variabili esplicative incluse nel dataset verranno trattate approfonditamente: i dettagli relativi ai dati mancanti per ogni variabile verranno esposti in questa sezione. Nessuna variabile esplicativa risulta avere un numero di valori mancanti eccessivamente elevato.

3.5 Variabili esplicative

Prima di cominciare con le analisi statistiche vere e proprie, è stata effettuata un'accurata analisi per ogni singola variabile esplicativa presente nel dataset. In questo paragrafo, queste variabili verranno esposte raggruppando per area di appartenenza, seguendo lo stesso criterio utilizzato nella sezione 2.3.

Nel dataset ci sono state fornite anche delle variabili esplicative relative ai cicli mestruali successivi al ritorno della fertilità ma, non essendo queste durate di interesse per il nostro studio, non sono state considerate, e non verranno trattate in questa sezione.

3.5.1 Variabili sociali

Nessuna delle cinque variabili relative alla sfera sociale delle donne in esame è stata inclusa nei diversi modelli statistici successivi.

Le tre variabili "Stato Civile", "Etnia" e "Religione" non sono state incluse nelle analisi statistiche per due motivi principali. Innanzitutto, perché la numerosità delle diverse classi risulta del tutto sbilanciata. Con l'eccezione di quattro valori mancanti, le donne risultano essere tutte sposate; tutte le donne sono di etnia caucasica, con l'eccezione di una donna di etnia asiatica, di una ispanica e di tre non meglio specificate; inoltre, solamente cinque soggetti non appartengono alla confessione religiosa dei Cristiani Cattolici. Inoltre, anche se le numerosità delle diverse classi fossero state ben bilanciate, queste variabili non sarebbero comunque state incluse nelle analisi statistiche: infatti, a rigor di logica, non sembrerebbe sensato assumere un qualche rapporto di causa/effetto o un qualche nesso di correlazione con le durate d'interesse.

Anche le due variabili "Anni di Matrimonio" e "Anni di Istruzione" non sono state incluse nelle analisi statistiche successive: anche per queste due quantità non sembrerebbe sensato assumere l'esistenza di una qualche relazione con la variabile d'interesse. Per controprova, queste due variabili sono state incluse nelle prime semplici analisi statistiche: non essendo mai stata rilevata alcuna significatività, abbiamo deciso di non includerle più in alcun modello.

3.5.2 Variabili fisiologiche

Come già osservato in studi precedenti (İnce *et al.*, 2007; Valeggia e Ellison, 2009), questo gruppo di variabili sembra avere un notevole effetto

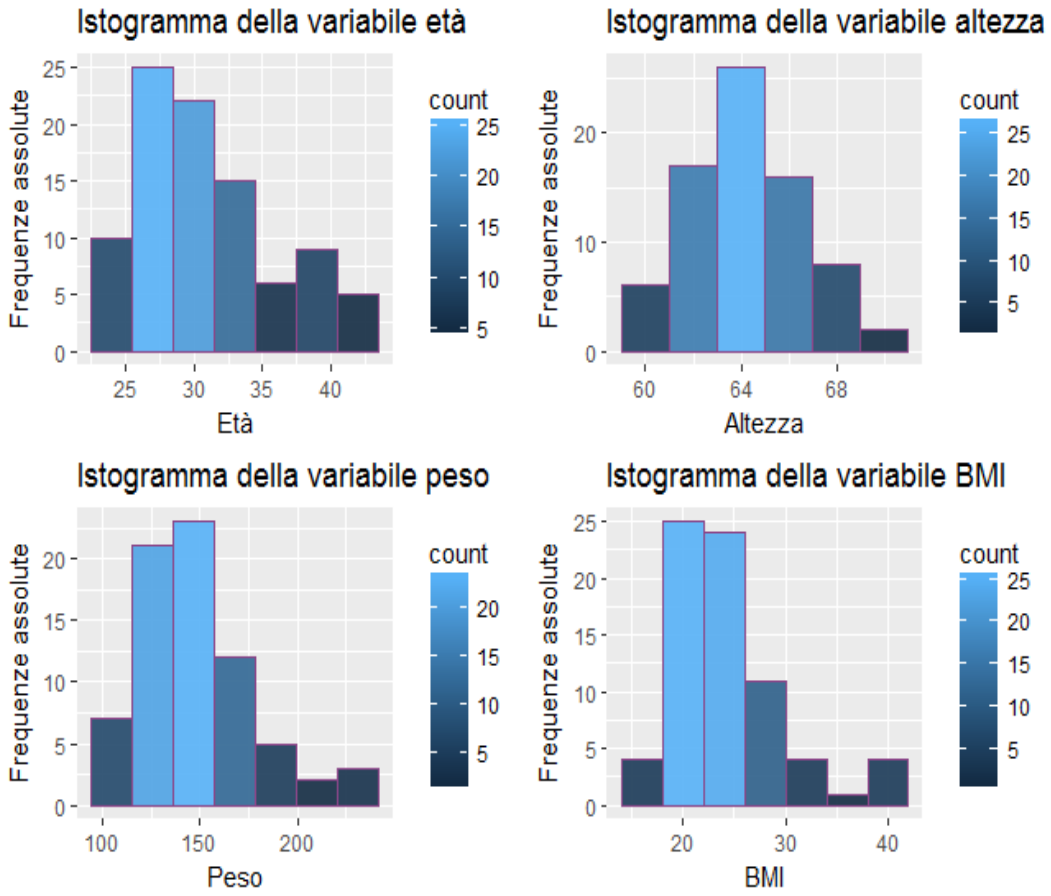


Figura 3.2: Istogrammi relativi alle variabili "Età", "Altezza", "Peso", "BMI"

sulla durata del periodo di amenorrea post parto.

Le quattro variabili relative alle caratteristiche fisiologiche delle donne ("Età", "Altezza", "Peso", "Indice di Massa Corporea") non sembrano comportare alcun tipo di problematica: osservando gli istogrammi relativi a queste quattro variabili, riportarti in figura 3.2, non notiamo nessun possibile comportamento anomalo causato da una densità troppo irregolare.

Le due variabili "Numero di Gravidanze" e "Figli Vivi" risultano fortemente correlate tra loro, come sembrerebbe essere logico: l'indice di correlazione empirico supera il 90%. Come verrà ribadito successivamente, la seconda variabile non risulterà significativa nelle nostre analisi, al netto della prima. [1]

Uno dei difetti del nostro dataset è dato dal fatto che il BMI sia

[1] Il numero di figli vivi per donna nel campione risulta piuttosto elevato. La media di figli nel campione è approssimativamente pari a 3, mentre più del 33% di donne nel campione ha più di tre figli!

espresso come variabile statica e non sotto forma di variabile dipendente dal tempo: nel già citato Valeggia e Ellison (2009) viene mostrato come nei modelli statistici adattati ai loro dati la variabile fissa BMI non risulti particolarmente significativa, mentre la variazione mensile di indice di massa corporea risulti molto informativa nel predire il ritorno alla fertilità. Nel nostro dataset è riportato solo il valore del BMI calcolato al momento dell'ingresso dei soggetti nello studio: ci accontenteremo di questa misura, tenendo ben presente che in uno studio successivo disporre dell'informazione relativa alle variazioni mensili di indice di massa corporea potrebbe risultare molto più informativo.

Le variabili relative all'età, al numero di gravidanze avute ed il numero di figli attualmente in vita hanno pochi dati mancanti, mentre le variabili relative all'altezza, al peso e al BMI delle donne risultano avere circa un 20% di informazioni mancanti.

3.5.3 Variabili relative ai biomarcatori ormonali

Nel dataset sono presenti numerose variabili ricavate dai valori dell'output dell'EHFM. Come già accennato nella sezione 2.3 abbiamo a nostra disposizione le date del primo alto e del primo picco registrato durante la fase di amenorrea post-parto ed il numero di bassi, alti e picchi rilevati durante questo periodo. Ci è stata inoltre fornita la data della stima del giorno di ovulazione (EDO), ottenuta prendendo come sua stima la data dell'ultimo picco osservato prima del ritorno alla fertilità: nel caso di un solo picco ormonale, la data dell'EDO sarà uguale alla data dell'unico picco oservato. Abbiamo inoltre incluso nel gruppo relativo ai biomarcatori ormonali le due variabili relative alla prima mestruazione post parto che segna la fine della fase di amenorrea: il fatto che le mestruazioni fossero fertili o meno e la somma dei valori di sanguinamento ottenuti durante questa fase.

Come già accennato in precedenza, il fatto di dipendere dal tempo e dalle durate di interesse rende queste variabili molto più delicate da trattare ed interpretare rispetto ai due gruppi di variabili descritte nei paragrafi precedenti.

Nelle colonne del dataset relative alle variabili dei biomarcatori ormonali, inoltre, sono presenti molti valori mancanti che, come già spiegato in precedenza, possono andare ad indicare due tipi di informazione mancante differente. Senza entrare nello specifico, circa il 30% dei valori di queste variabili risulta essere mancante.

A complicare ulteriormente le cose, si aggiunge il fatto che il monitoraggio elettronico che ci ha appunto permesso di ottenere queste informazioni è risultato pieno di incertezza e piuttosto incompleto. Al variare delle unità i momenti di inizio e di fine del monitoraggio elettronico risultano molto variabili; inoltre, non tutti i giorni sono stati monitorati. Questa insicurezza nella rilevazione di queste variabili rende i risultati delle analisi meno affidabili, sottraendo efficienza alle nostre stime ed aumentando il rischio di introdurre distorsione nelle nostre analisi.

Le variabili relative al numero di bassi, di alti e di picchi registrati durante il Ciclo 0 risultano essere pesantemente influenzate dalla durata del ciclo stesso: nelle analisi successive queste variabili saranno sempre inserite riscalandole rispetto ad un certo valore di tempo trascorso. Se non eseguissimo quest'operazione, le variabili relative al numero di valori ormonali monitorati si comporterebbero come variabili *leaker* per la variabile risposta: utilizzeremmo nelle nostre analisi statistiche variabili esplicative dipendenti dalla variabile risposta, cosa assolutamente errata concettualmente.

Per evitare questo fenomeno, il numero di bassi è stato diviso per il numero complessivo di giorni monitorati all'interno del Ciclo 0, il numero di alti per il numero di giorni compreso tra l'osservazione del primo alto ed il ritorno della fertilità, il numero di picchi per il numero di giorni compresi tra l'osservazione del primo picco e la fine del Ciclo 0. I valori ottenuti saranno poi moltiplicati per 7, in modo da ottenere una misura del numero medio di osservazioni settimanali: pur non essendo quest'operazione necessaria, è stata eseguita per ottenere dei valori meno prossimi allo 0, e rendere quindi di più facile interpretazione i coefficienti associati a queste tre variabili.

La medie relative al numero assoluto di bassi, alti e picchi registrati durante la fase di amenorrea post-parto sono pari rispettivamente a 78.97, 35.91 e 2.18; dopo aver riscalato queste variabili, otteniamo una media di 5 bassi settimanali, di 2.38 alti a settimana e di 1.77 picchi settimanali per il Ciclo 0. È bene notare come sotto a questa codifica ci sia un'assunzione piuttosto forte, e probabilmente errata, ossia che il numero di valori ormonali si distribuisca in maniera uniforme per tutta la durata del relativo ciclo. Naturalmente questo approccio può portare a distorsioni e perdita di efficienza ma, non avendo informazioni sulla forma almeno approssimata della tipica distribuzione dei valori relativi ai biomarcatori ormonali, abbiamo deciso di mantenere questa codifica nonostante siamo ben consci dei suoi limiti.

Nel Ciclo 0, tipicamente l'ordine dei tre eventi relativi ai valori or-

monali è il seguente: prima osserviamo il primo alto, poi il primo picco, infine il ritorno della fertilità vera e propria. Ci sono tuttavia numerosi casi in cui questa successione è violata. 13 casi accertati hanno una mestruazione senza che ci sia alcun segnale di picco ormonale, in tre casi il primo alto segnalato dell'EHFM avviene dopo il ritorno alla fertilità; ci sono altri tre casi, invece, in cui viene rilevato prima il primo picco ormonale del primo alto.

Nel dataset sono presenti anche due variabili relative al sanguinamento registrato alla fine del Ciclo 0: la prima è una variabile dicotomica che dice se la fine del Ciclo 0 sia stata demarcata o no da una mestruazione fertile, la seconda riporta il valore totale del punteggio di sanguinamento registrato. Non è stato possibile inserire nelle analisi relative al ritorno della fertilità queste due variabili, essendo avvenute temporalmente nello stesso momento in cui si è verificato l'evento di interesse.

3.5.4 Variabili metaboliche

Andiamo a vedere nel dettaglio le due variabili relative all'alimentazione del bambino: l'allattamento esclusivo al seno e l'introduzione dei "Supplementi". Essendo queste due variabili di forte interesse per i ricercatori, ed essendo state raccolte e successivamente codificate nel dataset in maniera tutt'altro che ottimale, verranno trattate in maniera approfondita nei due sottoparagrafi successivi.

Allattamento

Come già menzionato nel Capitolo 2, l'allattamento al seno esclusivo risulta essere un notevole fattore di allungamento dei tempi di ritorno alla fertilità. Questa peculiarità è alla base di molti protocolli per la prevenzione dopo il parto di gravidanze indesiderate: il cosiddetto *Lactation Amenorreah Method*, (Perez *et al.*, 1992), ad esempio, è un metodo particolarmente efficace per evitare il concepimento di ulteriori figli nei primi sei mesi dopo il parto: è stato dimostrato che le donne che in questo periodo allattano esclusivamente al seno e non hanno alcuna perdita di sangue hanno solo il 2% di probabilità di restare nuovamente incinte. (Kennedy *et al.*, 1989).

La variabile "Allattamento", suddivisa in due categorie, risulta essere sufficientemente bilanciata: nel campione sono presenti 53 donne che nel corso dell'analisi hanno mantenuto un regime di allattamento esclusivo

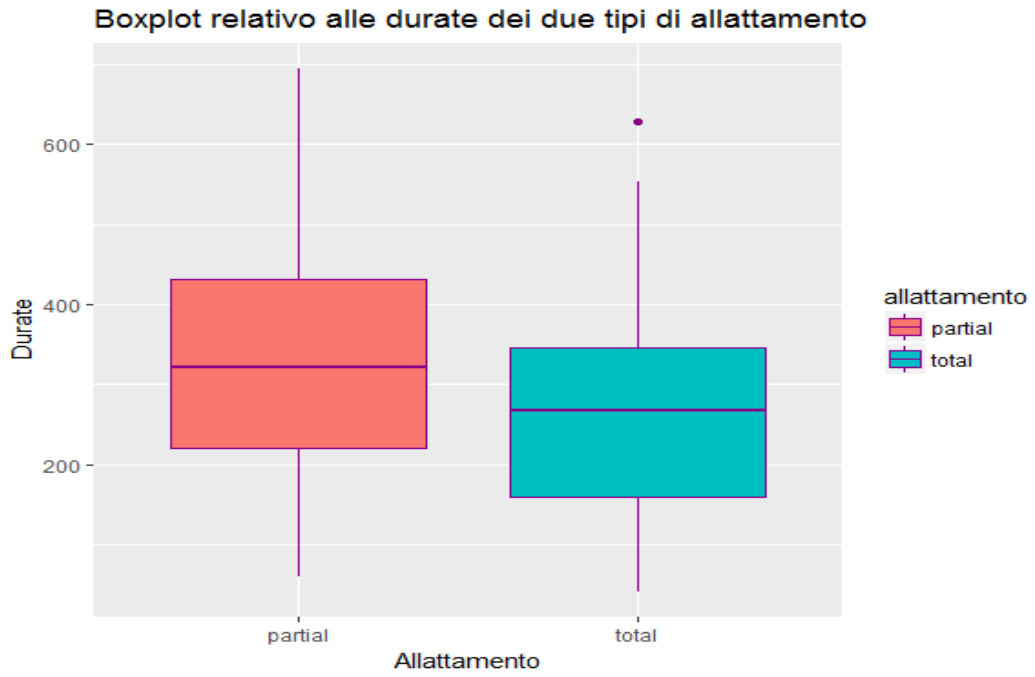


Figura 3.3: *Boxplot* della distribuzione dei tempi di ritorno alla fertilità rispetto ai due gruppi di allattamento.

al seno, mentre 39 soggetti hanno alimentato al seno solo parzialmente. Nel campione sono presenti solamente quattro valori mancanti relativi a questa variabile.

Per quanto questa variabile possa essere importante nel nostro tipo di studio, nel dataset è stata codificata mediante una semplice variabile dicotomica di livello "totale" o "parziale": se avessimo avuto a nostra disposizione una variabile dipendente dal tempo che ci avesse fornito la data di cessazione dell'allattamento esclusivo al seno, avremmo potuto valutare l'effetto di questo fattore con molta più efficacia. Inoltre, questa codifica ci porta ad avere un problema piuttosto importante: le unità che hanno la durata della fase di amenorrea molto breve hanno avuto poco tempo a loro disposizione per sperimentare l'evento "interruzione dell'allattamento totale". Il risultato è un forte effetto di selezione delle unità: sarà presente una forte distorsione verso l'alto nella stima del rischio di sperimentare la fine del Ciclo 0 per lo strato relativo all'allattamento totale. Questa distorsione si nota perfettamente nelle figure 3.3 e 3.4: in questi due grafici vengono riportati i *boxplot* relativi alle durate suddivise per i due gruppi di regime di allattamento e la stima stratificata della funzione di sopravvivenza, ottenuta mediante il noto stimatore di Kaplan-Meyer (Kaplan e Meier, 1958). Notiamo come la dimensione della distorsione introdotta sia tale da oscurare il reale effetto

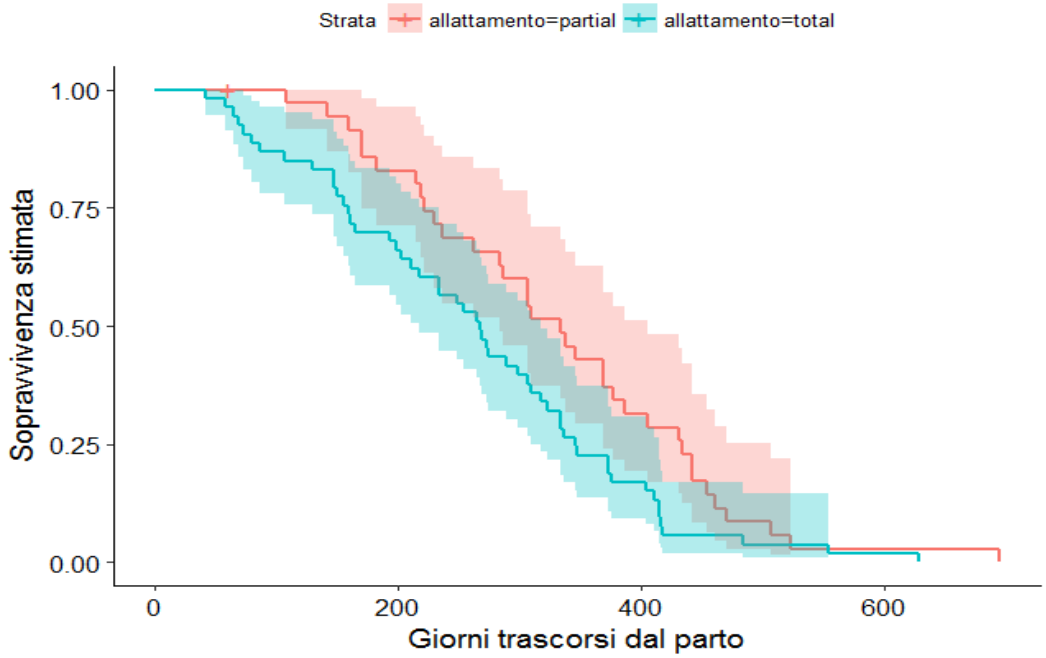


Figura 3.4: Stima di Kaplan-Meyer della funzione di sopravvivenza stratificata rispetto ai due gruppi di allattamento.

dell'interruzione dell'allattamento esclusivo al seno, e di farci sembrare che le donne che hanno interrotto questo protocollo sperimentino delle durate mediamente più lunghe (seppure gli intervalli di confidenza ottenuti dallo stimatore di Kaplan-Meyer siano quasi sempre intersecanti). Alla luce di tutti gli studi effettuati in precedenza sull'argomento, infatti, il fatto che le donne che hanno interrotto l'allattamento al seno esclusivo abbiano dei tempi di ritorno alla fertilità mediamente più lunghi risulta del tutto insensato.

Al crescere del tempo di permanenza nello studio, l'effetto di distorsione dovuto a questo *drop-out* di unità a rischio viene sempre di più mitigato: col tempo, le unità che non hanno avuto modo di aver già sperimentato l'evento "interruzione dell'allattamento totale" saranno già uscite dall'insieme dei soggetti a rischio, e le stime legate al rischio di ritorno della fertilità presenteranno sempre meno distorsione; questo fatto verrà riscontrato nei capitoli successivi, in cui verranno presentati i modelli statistici utilizzati per rispondere alle domande di interesse.

Supplementi

La seconda variabile relativa al regime alimentare del bambino è data dalla variabile "Supplementi". I valori di questa variabile indicano il

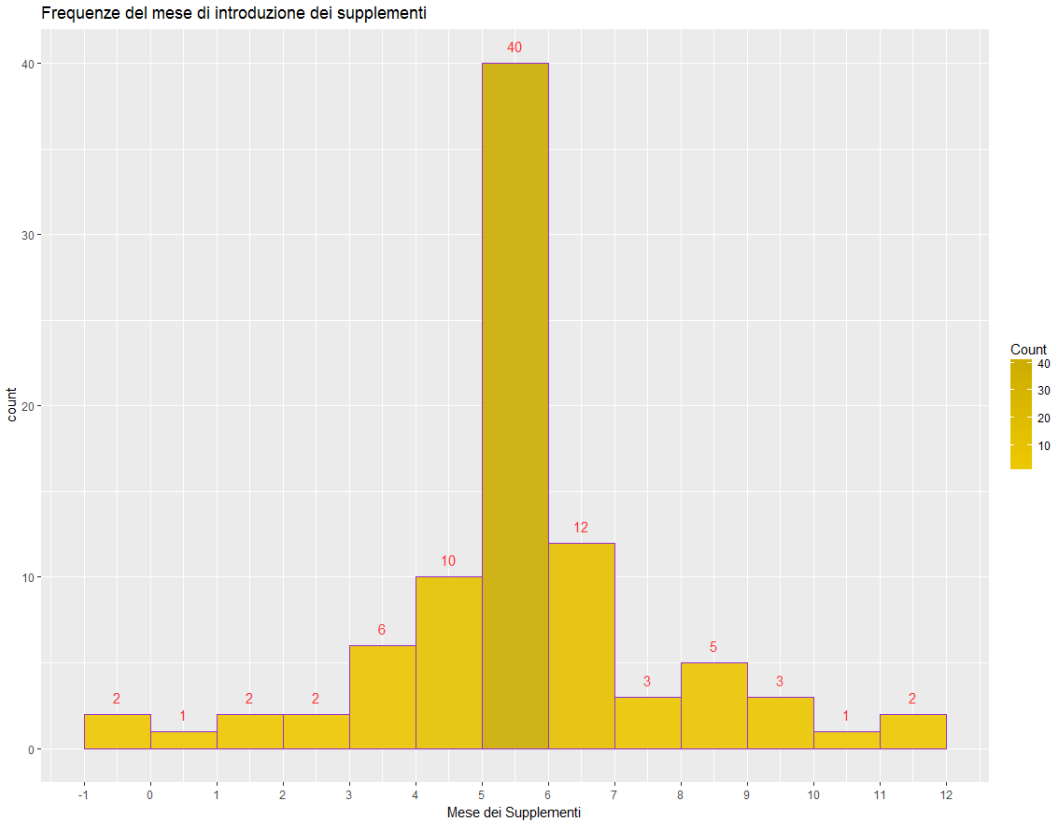


Figura 3.5: Distribuzione del mese di introduzione dei supplementi

mese in cui il bambino è stato "svezzato", ossia in cui la madre ha cominciato ad introdurre nella sua dieta alimenti sia solidi che liquidi diversi dall'esclusivo latte materno. I valori a nostra disposizione ci sono stati comunicati direttamente dalle donne appartenenti allo studio: purtroppo non è stato possibile reperire una data precisa per l'introduzione dei supplementi, ma solo il mese di età del bambino in cui questi sono stati introdotti. L'unità di misura temporale utilizzata nella codifica di questa variabile è stata modificata da mesi a giorni semplicemente moltiplicandone il valore per 30, senza porsi il problema della durata effettiva dei diversi mesi trascorsi. Non avendo dettagli sul giorno preciso in cui lo svezzamento è avvenuto, ma solo sul mese, questa approssimazione ci è sembrata del tutto sensata.

In figura 3.5 notiamo la distribuzione della variabile in questione: la moda del mese d'introduzione dei supplementi è pari a 6, il valore massimo è pari 12 mesi. Notiamo anche come due donne abbiano introdotto i supplementi già durante il primo mese di vita del lattante. Facendo un confronto con le durate relative al Ciclo 0, notiamo inoltre come più del 75% delle donne partecipanti allo studio abbia introdotto i supplementi prima del ritorno alla fertilità. Come ci è stato confer-

mato dai ricercatori, i valori mancanti codificati in questa variabile non stanno ad indicare che non sono stati forniti supplementi al bambino, ma semplicemente che non siamo riusciti a reperire questa informazione.

A prima vista sembra molto strano che questa variabile non rappresenti lo stesso identico fenomeno di quella precedente, ma, osservando per ogni donna i valori di ritorno alla fertilità e confrontandoli con queste due variabili, si è notato che in molti casi i valori non coincidessero: molte donne hanno tenuto un regime di allattamento totale al seno ma hanno dichiarato di aver cominciato l'utilizzo di supplementi in mesi precedenti al ritorno della fertilità, e viceversa. I ricercatori stessi ci hanno comunicato come, col senno di poi, la domanda non risultasse molto chiara nel questionario e che, forse, avrebbero dovuto essere più specifici.

Come per la variabile "Allattamento", anche questa variabile ha solo 4 valori mancanti che, tra l'altro, non coincidono tra di loro: possiamo interpretare questa discrepanza come un ulteriore indizio che ci fa propendere per il fatto che queste due variabili rappresentino effettivamente due entità diverse. Anche in questo caso seguiremo l'approccio adottato per molte altre variabili esplicative in questo studio: pur tenendo conto di non disporre di un'informazione ottimale per questa variabile, la inseriamo nei nostri modelli tenendo ben presente che i risultati potrebbero presentare una certa distorsione, o che potremmo non essere in grado di cogliere a pieno il reale effetto di questa variabile.

3.5.5 Variabili relative alla durata del monitoraggio elettronico

Nelle nostre analisi l'informazione relativa alla finestra del periodo di monitoraggio non è stata inclusa direttamente. Per tutte le unità statistiche, il monitoraggio ormonale comincia in un periodo compreso tra la nascita del figlio ed il giorno del ritorno alla fertilità, mentre termina sempre dopo la fine della fase di amenorrea post parto. Nel Ciclo 0 abbiamo notato una perfetta coincidenza tra l'avere attivo il monitoraggio ormonale e l'avere informazioni relative ai biomarcatori ormonali. Abbiamo a disposizione il numero totale di giorni monitorati elettronicamente durante il periodo di permanenza nello studio per ogni unità, tuttavia non abbiamo informazioni aggiuntive sul quando i giorni non monitorati siano stati tralasciati.

Nonostante la scelta di non includere esplicitamente nel modello variabili relative alla durata del monitoraggio elettronico, abbiamo co-

munque adottato alcune soluzioni per cercare di mitigare l'incertezza causata dal fatto di non riuscire ad osservare l'esito del monitoraggio ormonale per ogni singolo giorno di studio. Riscalando le variabili relative al numero di bassi/alti/picchi mediante il procedimento descritto nella sezione 3.5.3, oltre che a risolvere il problema di *leakage* discusso in precedenza, dovremmo riuscire ad ottenere delle variabili i cui valori siano meno sensibili al numero di giorni che non siamo riusciti a monitorare.

Come molti altri aspetti di questa analisi, è bene tenere presente che se avessimo avuto dei dati raccolti con maggiore precisione, avremmo potuto ottenere da questo studio dei risultati più attendibili; non essendo così, cercheremo di effettuare le analisi al meglio, tenendo presente che i risultati ottenuti dalle nostre analisi potrebbero non essere eccessivamente affidabili, essendo stati ottenuti da un dataset con una numerosità campionaria non eccessivamente elevata, un certo numero di valori mancanti e con variabili raccolte in maniera non ottimale.

Capitolo 4

Ritorno della fertilità dopo il parto: analisi statistiche

In questo capitolo verranno presentate le analisi statistiche effettuate per rispondere al primo dei quesiti riportati nella sezione 2.4. È di interesse comprendere quali siano i fattori che influenzano il "rischio" di ritorno della fertilità dopo il parto, ed in quale dimensione. Verrano presentati i modelli statistici adattati ai dati ed i risultati ottenuti. Le analisi statistiche sono state svolte mediante il linguaggio di programmazione *R* (R Core Team, 2016).

4.1 Le variabili selezionate per l'analisi

Dal dataset originale, abbiamo dovuto operare una selezione di variabili da inserire nei modelli relativi all'analisi di interesse. La variabile risposta in questi modelli è la durata complessiva del Ciclo 0, ottenuta mediante la procedura descritta nella sezione 3.2. La variabile relativa alle censure è stata ottenuta mediante la procedura descritta nella sezione 3.3; come già fatto osservare, per il Ciclo 0 abbiamo un solo caso censurato.

Le variabili fisiologiche che abbiamo considerato sono età, altezza, peso e indice di massa corporea. Come variabili relative al monitoraggio elettronico abbiamo incluso il numero di bassi, alti e picchi rilevati durante la fase di amenorrea, la data del primo picco e la data del primo alto. Abbiamo inoltre incluso il numero di gravidanze avute ed il numero di figli attualmente in vita, insieme alle due variabili relative all'alimen-

tazione del bambino: l'allattamento ed il numero di giorni compresi tra il parto e l'introduzione dei supplementi. Non abbiamo inserito la variabile "EDO" poiché risulterebbe essere una variabile *leaker* della risposta: essendo questa definita come l'ultimo picco prima della fine del ciclo, abbiamo che il suo valore dipende dall'esito della variabile risposta, non risulta quindi sensato utilizzarla come variabile regressiva. Pur essendo consci dell'insensatezza di questa operazione, abbiamo comunque provato ad inserirla nei primi modelli adattati: come ci si aspettava, il suo potere predittivo è risultato molto elevato.

Nessuna variabile relativa alla sfera sociale è stata inclusa nei modelli.

In questo dataset sono presenti sia variabili statiche, ossia il cui valore non cambia con lo scorrere del tempo, che variabili dipendenti dal tempo, il cui valore si modifica in funzione del tempo. Il fatto di avere variabili dipendenti dal tempo nell'ambito di dati di sopravvivenza rende la analisi statistiche decisamente più complesse, escludendo inoltre automaticamente la possibilità di adattare modelli che non siano in grado di trattare questo particolare tipo di variabili.

Le variabili che in questo dataset verranno trattate come dipendenti dal tempo sono il primo alto registrato, il primo picco osservato e l'introduzione dei supplementi. Le variabili relative al numero di bassi, alti e picchi monitorati nel Ciclo 0 sarebbero potute essere trattate come dipendenti dal tempo, potenzialmente, ma nella codifica in cui ci sono state fornite possono essere trattate unicamente come variabili statiche. Come già ribadito nella sezione 3.5, sarebbe stato più appropriato avere le due variabili relative all'indice di massa corporea ed all'interruzione dell'allattamento esclusivo al seno espresse come variabili dipendenti dal tempo. Non essendo possibile risalire a questa informazione, svolgeremo le analisi semplicemente con le variabili che abbiamo a disposizione.

4.2 Il modello di Cox a rischi proporzionali

In questa sezione presenteremo brevemente uno dei modelli più utilizzati nelle analisi statistiche nell'ambito dei dati di sopravvivenza, il modello semiparametrico di Cox a rischi proporzionali (Cox, 1972). La conoscenza di questo particolare modello statistico sarà di fondamentale importanza per comprendere le analisi svolte in questa tesi, sia come modello a se stante che come punto di partenza per modelli più elaborati utilizzati nel Capitolo 5.

La funzione di rischio è una delle funzioni più importanti nell'ambito dei dati di durata. Questa misura il tasso istantaneo di transizione di stato al tempo t per la variabile casuale T:

$$\lambda(t) = \lim_{\Delta T \to 0} \frac{P(t \leq T < t + \Delta T | T \geq t)}{\Delta T} \tag{4.1}$$

La funzione di rischio è non negativa per definizione.

Il modello di Cox a rischi proporzionali modella la funzione di rischio $\lambda_i(t)$ per un particolare soggetto i, i=1, ..., n, mediante la seguente formulazione:

$$\lambda_i(t) = \lambda_0(t) g(\mathbf{x}_i \boldsymbol{\beta}), \tag{4.2}$$

con $\lambda_0(\cdot)$ funzione del tempo t che rappresenta il rischio di base comune a tutte le unità presenti nel campione, $\mathbf{x}_i$ vettore di dimensione $1 \times p$ delle covariate relative all'i-esima unità statistica e β vettore di dimensione $p \times 1$ dei relativi coefficienti. La funzione $g(\cdot)$ è scelta come funzione con supporto $\mathbb{R}^+$, dato che la funzione di rischio per definizione non può essere negativa. In genere, si utilizza la funzione esponenziale: poniamo quindi $g(\cdot) = exp(\cdot)$. Il modello di Cox è un modello di tipo semiparametrico perché composto da due termini: λ_0, la cosiddetta *baseline*, la cui stima è ottenuta in maniera non parametrica ed il termine esponenziale, la parte parametrica, che dipende dai parametri relativi alle covariate.

Il modello a rischi proporzionali è così chiamato perché il rapporto tra le funzioni di rischio di due soggetti è una proporzione fissa costante nel tempo. Infatti:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) e^{\mathbf{x}_i \boldsymbol{\beta}}}{\lambda_0(t) e^{\mathbf{x}_j \boldsymbol{\beta}}} = e^{(x_{1i} - x_{1j})\beta_1 + \ldots + (x_{pi} - x_{pj})\beta_p}. \tag{4.3}$$

Il modello di Cox tratta senza problemi sia variabili continue che variabili categoriali: per ogni variabile categoriale con K classi saranno presenti $K - 1$ parametri nella parte parametrica del modello; l'effetto della classe rimanente sarà invece incluso nella stima non parametrica del rischio di base.

Il metodo per ottenere le stime del vettore dei coefficienti β si basa sul massimizzare la funzione di Verosimiglianza Parziale (Cox, 1975),

ottenuta a partire dalla funzione di rischio. Tipicamente, questa operazione viene svolta mediante un algoritmo iterativo come, ad esempio, il metodo di Newton-Rhapson. Definiamo C_i come l'indicatore di censura per l'i-esimo evento, che assume valore 1 per i casi completi e 0 per i casi censurati, ed $R(t)$ come l'insieme dei soggetti a rischio di sperimentare l'evento al tempo t, ossia l'insieme delle unità che non hanno ancora sperimentato l'evento d'interesse e che non sono uscite dallo studio per altri motivi, andando così a creare una durata censurata. Nella formula 4.4 viene espressa la funzione di verosimiglianza parziale:

$$L(\beta) = \prod_{i:C_i=1} \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{\sum_{j \in R(t)} e^{\mathbf{x}_j \boldsymbol{\beta}}} \tag{4.4}$$

La stima del valore dei coefficienti non dipende dalla stima della funzione di rischio di base, che può essere ottenuta non parametricamente mediante lo stimatore di Breslow (Breslow, 1974; Hanley, 2008). È possibile attribuire alla funzione di *baseline* il significato di stima della funzione di rischio per un soggetto (anche ipotetico) che abbia tutte le covariate continue di valore pari a 0 e tutte le covariate categoriali con valore pari alla classe di riferimento. Per questo motivo, spesso si sceglie di prendere come livello di base per le variabili categoriali la classe più frequente, e di riscalare le variabili continue in modo da poter attribuire alla baseline un significato sensato (ad esempio, sottraendo ad ogni variabile continua la propria media, è possibile ottenere una funzione di rischio di base per "l'individuo medio").

L'interpretazione dei parametri associati alle diverse variabili esplicative è la seguente: parametri positivi causano un aumento della funzione di rischio al crescere del valore della relativa covariata, al netto delle altre variabili; al contrario, parametri negativi riducono il rischio. Nella sua formulazione originale, il modello di Cox prevede che nello stesso punto temporale non possano verificarsi più eventi contemporaneamente: sono presenti numerosi metodi per aggiustare la verosimiglianza parziale in modo da poter trattare gli eventi avvenuti simultaneamente (i cosiddetti dati *"tied"*) (Therneau e Grambsch, 2013).

Il modello a rischi proporzionali è un modello piuttosto semplice ma incredibilmente versatile: numerose estensioni sono state introdotte per superarne i limiti (Therneau e Grambsch, 2013). Ad esempio, è possibile stimare diverse funzioni di rischio di base per diversi strati, è possibile inserire un termine di interazione tra il tempo e le variabili, oppure si

può fare in modo che i coefficienti dipendano da una qualche funzione del tempo.

Il modello di Cox stratificato è una delle estensioni più popolari del modello a rischi proporzionali: data una variabile categoriale con K classi è possibile stimare K diverse funzioni di rischio di base, una per ogni modalità. La formulazione del rischio per l'i-esima unità la cui variabile di stratificazione assume valore k diventa quindi:

$$\lambda_i(t) = \lambda_{k0}(t)e^{\mathbf{x}_i\boldsymbol{\beta}}, \tag{4.5}$$

Questa particolare estensione è tipicamente utilizzata per trattare il fatto che una variabile violi l'assunzione di proporzionalità dei rischi: per risolvere questo problema è sufficiente stimare K *baseline* diverse. Nel caso di variabili continue, è prima necessario discretizzarle in un certo numero di classi per adattare un modello stratificato. È inoltre possibile ottenere dei parametri specifici per le diverse classi della variabile stratificata inserendo dei termini di interazione tra questa e le variabili esplicative di cui si vuole valutare l'effetto al variare degli strati.

Il modello di Cox risulta particolarmente indicato per trattare variabili dipendenti dal tempo, grazie alla particolare formulazione della funzione di verosimiglianza parziale. Uno dei metodi più semplici ma più efficaci per trattare variabili esplicative dipendenti dal tempo con valori costanti a tratti è il seguente: è sufficiente suddividere l'i-esima osservazione in k intervalli temporali, introducendo un nuovo intervallo ogni volta che si osserva una variazione in almeno una delle variabili dipendenti dal tempo. L'informazione relativa alle durate viene codificata mediante due diverse variabili indicanti il tempo di inizio ed il tempo di fine dell'intervallo. Tutti gli intervalli eccetto l'ultimo vengono considerati come casi censurati, in modo da non creare eventi fittizi. Per come sono stati formulati gli estremi degli intervalli, un solo intervallo per volta tra quelli relativi allo stesso soggetto può essere presente nell'insieme $R(t)$ degli individui a rischio.

Il fatto che il modello a rischi proporzionali sia in grado di trattare variabili dipendenti dal tempo risulterà particolarmente utile nelle nostre analisi, dato che le variabili dipendenti dal tempo presenti nel nostro dataset risultano essere costanti a tratti.

4.3 Imputazione dei dati mancanti

Come già spiegato nella sezione 3.4, nel dataset relativo alla nostra analisi sono presenti numerosi valori mancanti. Il problema dei dati

mancanti in statistica è un problema ben noto, ed è stato ampiamente trattato nella letteratura specifica, si veda ad esempio Schafer e Graham (2002). Senza addentrarsi troppo nei dettagli, esistono numerosi approcci per trattare questa particolare problematica. Tipicamente, il fatto di avere dei valori mancanti all'interno di un insieme di dati da analizzare rende le analisi statistiche più complesse da svolgere, ed introduce ulteriore incertezza nei risultati. Non sono molti i modelli statistici in grado di gestire automaticamente i valori mancanti: l'approccio di *default* di diversi *software* statistici è quello di eliminare direttamente dal dataset le righe relative alle unità statistiche con almeno un dato mancante e di adattare il modello in questione sull'insieme ridotto dei dati così ottenuto. Questo approccio presenta due principali difetti: innanzitutto, il fatto di eliminare delle unità statistiche comporta una perdita di efficienza che può essere anche molto pesate nel caso in cui la numerosità del campione non sia elevata o il numero di soggetti scartati risulti particolarmente alto. Inoltre, c'è il rischio di introdurre distorsione nelle stime nel caso in cui la distribuzione della variabile casuale $M \in \{0, 1\}$, indicatrice del fatto che un dato sia mancante o meno, dipenda in qualche modo dalle caratteristiche degli individui (Schafer e Graham, 2002; van der Heijden *et al.*, 2006). Le tipologie di dati mancanti si possono raggruppare in tre principali gruppi, a seconda delle caratteristiche relative alla distribuzione della variabile M:

- *Missing Completely at Random (MCAR)*: la distribuzione di M è completamente casuale, e non dipende dai valori relativi all'unità statistica;
- *Missing at Random (MAR)*: la distribuzione di M dipende unicamente dai valori osservati dei soggetti;
- *Missing Not at Random (MNAR)*: la distribuzione di M dipende sia dalle caratteristiche osservate del soggetto che da quelle mancanti.

Nel caso in cui il la distribuzione di M sia MCAR, il fatto di eliminare le righe con dati mancanti introduce semplicemente una perdita di efficienza. Se invece i dati sono MAR (o peggio, MNAR), utilizzare il metodo di cancellazione delle righe comporterà molto probabilmente l'introduzione di una qualche forma di distorsione sulle stime. Nel nostro caso, il metodo di eliminazione delle righe con almeno un dato mancante non è fattibile: eliminando queste unità avremmo una numerosità campionaria troppo bassa per effettuare delle analisi statistiche

affidabili. Resterebbero infatti a nostra disposizione solamente 38 unità statistiche delle 93 osservazioni presenti nel campione originario, numerosità del tutto inadeguata per adattare un modello statistico con 13 variabili esplicative. Inoltre, i nostri dati non sembrano essere MCAR: adattando per ogni variabile esplicativa con dati mancanti un modello di regressione logistica che utilizzi come variabile risposta il fatto di essere mancante e come covariate le rimanenti variabili del dataset, notiamo delle significatività associate a diversi coefficienti, a conferma del fatto che il meccanismo di censura dipenda (almeno) dai valori osservati delle rimanenti variabili (Garson, 2015).

Al fine di evitare questa patologica perdita di efficienza, sarà necessario effettuare una procedura di imputazione dei dati mancanti: ai valori mancanti presenti nel dataset vengono sostituiti dei valori "plausibili", ottenuti sulla base dell'informazione fornita dai dati osservati.

Esistono svariati metodi di imputazione dei dati. Uno degli approcci più immediati (considerato piuttosto *naïve*) prevede ad esempio di assegnare ai valori mancanti di una certa variabile la media dei suoi valori osservati. Un altro metodo molto semplice prevede l'introduzione di una nuova variabile dicotomica per ogni colonna in cui siano presenti dei dati mancanti, che associa ai valori completi il valore 1 e 0 all'eventualità di dati mancanti (Huberman e Langholz, 1999). I valori mancanti vengono quindi sostituiti nella variabile originale con valori arbitrari che risulteranno del tutto ininfluenti, essendo che l'informazione che apportano al modello statistico viene completamente colta dalla variabile dicotomica. L'utilizzo di questi metodi è sconsigliato nella maggior parte dei casi, in quanto spesso introducono perdita di efficienza e, nel caso in cui i dati siano MAR o MNAR, possono causare distorsioni anche non indifferenti (van der Heijden *et al.*, 2006).

Esistono metodi di imputazione dei dati molto più raffinati: una delle tipologie più popolari e che funziona meglio nei casi reali è l'Imputazione Multipla. Questo particolare approccio, introdotto in Rubin e Schenker (1986), ha ottenuto particolare successo grazie alle buone proprietà statistiche di cui gode. L'idea sottostante è quella di effettuare m imputazioni diverse, ottenendo così m dataset differenti, su cui lo stesso modello sarà adattato m volte. I valori dei parametri coi relativi errori standard verranno ottenuti mediante regole ben precise: così facendo, si riesce ad effettuare l'inferenza statistica tenendo conto anche della variabilità introdotta dalle stime imputate. Un metodo di imputazione multipla molto popolare ed efficace è denominato *"MICE"* (*"Multiple Imputation by Chained Equations"*, Van Buuren e Oudshoorn, 1999),

si veda ad esempio White *et al.* (2011) per i dettagli.

Altri metodi molto popolari sono basati invece sul concetto di Imputazione Singola: in questi metodi, l'obiettivo è quello di ricreare un dataset il più simile possibile all'insieme di dati che avremmo avuto a disposizione nel caso in cui fossimo riusciti ad osservare tutti i valori completi, in modo da poter applicare su questo dataset diversi procedimenti statistici di analisi come se avessimo a nostra disposizione un dataset senza alcun valore mancante. Alcuni esempi di metodi basati su questo principio si ritrovano in Andridge e Little (2010), Troyanskaya *et al.* (2001), Städler e Bühlmann (2010).

Il metodo utilizzato in questa tesi per effettuare l'imputazione dei dati mancanti nel nostro dataset appartiene a quest'ultima categoria, è denominato *missForest*, introdotto in Stekhoven e Bühlmann (2012). Questo particolare algoritmo di imputazione si basa sul noto modello *Random Forest* (Breiman, 2001): l'idea è quella di ottenere dei valori imputati per i diversi dati mancanti adattando iterativamente modelli di Foresta Casuale alle variabili del dataset.

Le funzioni che implementano l'algoritmo di imputazione sopracitato utilizzate per effettuare le nostre analisi sono contenute nel pacchetto R `missForest` (Stekhoven, 2015), disponibile su *CRAN*.

Il procedimento, data una matrice $\mathbf{X}$ di dimensione $n \times p$, è il seguente: innanzitutto, si riordinano le colonne della matrice sulla base della percentuale di dati mancanti, in ordine crescente; si effettua quindi una semplice imputazione *naïve* su tutti i valori mancanti del dataset, ad esempio mediante il metodo di imputazione del valore medio. Per ogni variabile s che presenta dati mancanti, definiamo $\mathbf{y}_{obs}^{(s)}$ come il vettore dei suoi valori completi, e $\mathbf{y}_{mis}^{(s)}$ come il vettore dei valori mancanti. Definiamo inoltre $\mathbf{i}_{obs}^{(s)} \subset \{1, \cdots, n\}$ come il vettore di dimensione $q \times 1$, $q<n$, degli indici associati alle righe della matrice $\mathbf{X}$ relative ai casi completi per la variabile s, $\mathbf{i}_{mis}^{(s)}$ invece come il vettore degli indici relativi ai valori mancanti.

Adattiamo quindi una *Random Forest* utilizzando come variabile risposta il vettore dei valori completi della variabile s e come variabili esplicative la matrice $\mathbf{X}_{obs}^{(s)}$ che ha come colonne tutte le variabili presenti nel dataset meno la variabile s e come righe esclusivamente le unità di indice $\mathbf{i}_{obs}^{(s)}$. Il passo successivo consiste nell'ottenere una nuova stima di $\mathbf{y}_{mis}^{(s)}$ sulla base dei valori predetti dalla *Random Forest*, utilizzando come insieme delle variabili esplicative la matrice $\mathbf{X}_{mis}^{(s)}$ definita, similarmente ad $\mathbf{X}_{obs}^{(s)}$, come la matrice che ha come colonne tutte le variabili

Algoritmo 1 Algoritmo di imputazione `missForest`

Richiede X matrice $n \times p$, γ criterio di arresto;
k vettore degli indici delle colonne con dati da imputare, ordinato per percentuale crescente di valori mancanti;
while γ non soddisfatta **do**
    $\mathbf{X}_{imp}^{old}$ assegna matrice imputata precedentemente
    for s in k **do**
        adatta una *Random Forest*: $\mathbf{y}_{\mathbf{obs}}^{(\mathbf{s})} \sim \mathbf{X}_{\mathbf{obs}}^{(\mathbf{s})}$
        predici $\mathbf{y}_{mis}^{(s)}$ usando $\mathbf{X}_{mis}^{(s)}$
        $\mathbf{X}_{imp}^{new}$ aggiornata utilizzando $\mathbf{y}_{mis}^{(s)}$;
    end for
    aggiorna γ
end while
Ritorna la matrice imputata $\mathbf{X}^{imp}$

tranne s, e come righe tutte le righe di indice $\mathbf{i}_{mis}^{(s)}$. Questo procedimento viene eseguito per tutte le colonne con dati mancanti, e viene iterato un numero N di volte finché non viene soddisfatto un certo criterio di convergenza γ.

Nell'algoritmo 1, il procedimento è spiegato in maniera più dettagliata.

Adesso esporremo i motivi per cui abbiamo scelto l'utilizzo del metodo di imputazione *missForest* rispetto ad un approccio di tipo MICE. Teoricamente, l'imputazione multipla ha proprietà statistiche più desiderabili rispetto ai metodi di imputazione singola: il fatto di tenere conto della variabilità introdotta dallo stimare modelli statistici utilizzando valori imputati è un approccio più appropriato nel fare inferenza (Royston *et al.*, 2004). Entrambi i metodi provocano tipicamente molta meno distorsione nelle stime dei parametri rispetto ai metodi *naïve*. Seppur nel caso di dati MCAR *missForest* sembri produrre meno distorsione rispetto a MICE (Waljee *et al.*, 2013), in presenza di dati MAR l'Imputazione Multipla non sembra produrre *bias*, al contrario dell'imputazione *missForest* (Shah *et al.*, 2014), che sembra causare una leggera distorsione.

Nel nostro caso specifico, tuttavia, abbiamo scelto di utilizzare un approccio di Imputazione Singola per diversi motivi. Innanzitutto, abbiamo bisogno di usare lo stesso dataset per adattare modelli anche molto diversi tra loro: per effettuare un'imputazione MICE è infatti necessario specificare il modello parametrico sul quale verranno adattati

i dati (il cosiddetto "modello sostanziale"). Senza porre questo vincolo, non è possibile ottenere dei valori imputati. Inoltre, adattare ai dataset imputati mediante un certo modello sostanziale un modello differente non è corretto: l'approccio MICE perde la caratteristica di tenere conto della variabilità dovuta all'imputazione. *MissForest*, invece, non ha bisogno di questa particolare specificazione ed il dataset ottenuto da questo tipo di imputazione può essere utilizzato per adattare modelli differenti. Il semplice fatto di avere un singolo dataset risulta inoltre molto più semplice ed immediato.

MICE, inoltre, ottiene stime dei parametri e relativi errori standard sulla base della cosiddetta regola di Rubin (Rubin e Schenker, 1986). Se il modello statistico utilizzato è di tipo non parametrico, in cui non ci siano parametri e relativi errori standard da stimare, l'approccio di stimare m modelli e poi combinarne i risultati perde di significato.

L'approccio MICE prevede l'associazione di una particolare distribuzione parametrica ad ogni variabile i cui valori debbano essere imputati; è possibile tuttavia utilizzare anche metodi non parametrici, come ad esempio, in Shah *et al.* (2014), in cui i valori da imputare nei diversi dataset vengono ottenuti mediante un approccio basato sulle Foreste Casuali, denominato *MICE Random Forest*. Abbiamo provato ad adattare un modello a rischi proporzionali ai dati utilizzando come metodo di imputazione il metodo sopracitato. I risultati ottenuti sembrano molto più conservativi rispetto alle stime ottenute utilizzando *missForest* come algoritmo di imputazione: gli errori standard associati ai singoli parametri risultano più elevati rispetto a quelli ottenuti in precedenza, ed i grafici diagnostici sembrano peggiori. Probabilmente in un problema come il nostro con numerosità campionaria ridotta può rivelarsi migliore un approccio più aggressivo (a costo di introdurre una possibile ma leggera distorsione delle stime) rispetto ad un metodo più conservativo come MICE, che al crescere della numerosità campionaria presenta distorsioni più basse, ma per numerosità poco elevate risulta essere meno efficiente.

4.4 Modello statistico adattato e risultati ottenuti

Come già accennato nella sezione 4.2, il modello statistico utilizzato per rispondere alla prima delle tre domande di interesse di questa analisi è il modello di Cox a rischi proporzionali. Il buon adattamento del modello ai dati è riscontrato mediante grafici e test diagnostici, che verrano

esposti nella sezione 4.4.1. Il dataset utilizzato è composto dalle variabili "Età", "Altezza", "Peso", "BMI", "Numero di Gravidanze", "Figli Attualmente in Vita", "Allattamento totale", "Numero di Bassi", "Numero di Alti", "Numero di Picchi", "Supplementi", "Primo Alto", "Primo Picco". Le tre variabili relative ai giorni di introduzione dei supplementi, di rilevazione del primo alto e di osservazione del primo picco sono dipendenti dal tempo: inizialmente sono state codificate sotto forma di variabile numerica che indica il numero di giorni trascorsi tra l'inizio del Ciclo 0 e la data dell'evento in questione; dopo l'operazione di suddivisione in sotto-episodi descritta nella sezione 4.2 sono state codificate sotto forma di variabili dicotomiche, con valore 1 nel caso in cui l'evento d'interesse sia avvenuto alla fine di uno dei sotto-episodi precedenti e valore 0 in caso contrario.

Prima di adattare il modello ai dati, è necessario effettuare l'imputazione dei dati mancanti. Nel nostro caso abbiamo utilizzato l'algoritmo *missForest*, presentato nel dettaglio nella sezione precedente. Come ogni metodo basato sulle Foreste Casuali, anche per *missForest* è necessario fornire alcuni parametri di regolazione, come il numero di alberi di cui comporre il nostro modello ed il numero di variabili da estrarre casualmente ad ogni nodo per effettuarne la divisione. Nel nostro caso abbiamo scelto di adattare foreste da 500 alberi, e di selezionare 7 variabili per ogni nodo. Questa scelta è stata fatta in maniera euristica ripetendo l'imputazione diverse volte e tenendo conto dei diversi errori *out-of-bag* (OOB) prodotti dalle diverse imputazioni. I due parametri sono stati quindi scelti sulla base dell'errore OOB che, mediamente, risultava più basso. Dopo aver fissato i due valori ed aver ottenuto il dataset imputato definitivo, abbiamo effettuato le operazioni di suddivisone del dataset in sotto-episodi, trasformando le tre variabili dipendenti dal tempo in variabili dicotomiche. Abbiamo quindi adattato il modello a rischi proporzionali e valutato sulla base dei grafici diagnostici e dell'AIC (*"Akaike Information Criterion"*, Akaike, 1974) quali variabili inserire nel modello finale. Una volta fissati i due parametri di regolazione dell'algoritmo di imputazione, tutte le operazioni descritte in precedenza sono state eseguite diverse volte per valutarne la robustezza: i risultati ottenuti dalle diverse imputazioni sono variati molto poco.

Uno dei nostri dubbi riguardo l'utilizzo di un metodo di imputazione dei dati mancanti è la possibilità di imputare valori relativi alle date del primo picco e del primo alto nei casi in cui effettivamente questi eventi non si siano verificati. Fortunatamente, nei casi in cui abbiamo la certezza che non ci siano stati alti e picchi ormonali nel Ciclo 0, abbiamo

osservato che i valori mancanti del primo alto e del primo picco sono stati sostituiti da valori più alti della durata stessa del Ciclo: sembrerebbe quindi che l'algoritmo di imputazione riesca a cogliere bene il fatto che nel Ciclo 0 non si siano verificati alti e picchi ormonali.

4.4.1 Dettagli e diagnostiche

Il modello di Cox a rischi proporzionali è stato adattato al dataset descritto nella sezione precedente, coi dati mancanti imputati. Per ogni variabile, abbiamo provato ad inserire i termini quadratici, ad effettuare trasformazioni logaritmiche, o ad inserirle sotto forma di radice quadrata. Si è provato inoltre ad inserire tutte le possibili interazioni di secondo livello tra le variabili considerate. Naturalmente, non è stato possibile adattare il modello che includesse sia le variabili esplicative che tutti questi termini quadratici e tutte le interazioni di secondo livello a causa della scarsa numerosità campionaria e del loro elevato numero. Abbiamo quindi provato ad inserire un termine quadratico o un'interazione alla volta: abbiamo annotato quelli che sono risultati almeno debolmente significativi ($p \leq 0.10$) e man mano li abbiamo inclusi nel modello completo finale, rimuovendo di volta in volta i termini che non risultavano significativi e non portavano un miglioramento in termini di AIC.

È stata infine eseguita una procedura di *stepwise backward* basata sull'indice AIC sul modello con tutte le variabili ed i termini quadratici e di interazione di secondo livello selezionati precedentemente, al fine di selezionare quali variabili ed interazioni fossero realmente influenti.

Nella tabella 4.1 sono riportati i valori dei diversi coefficienti associati alle variabili esplicative, insieme ai relativi errori standard, ai valori delle statistiche Z coi p-value associati. Sono anche riportati gli intervalli di confidenza di livello 95% relativi ai coefficienti I coefficienti associati ad alcune variabili non sono risultati significativi, tuttavia sono stati comunque inclusi nel modello finale sulla base della selezione effettuata dalla procedura di *stepwise* con AIC.

Il modello sembra adattarsi molto bene ai dati. Un grafico diagnostico molto utilizzato per verificare l'adeguatezza del modello di Cox si basa sulla distribuzione dei residui di *Cox-Snell* (Kay, 1977). Nel caso di corretta specificazione del modello, questi residui si distribuiscono come un campione censurato ottenuto da una distribuzione Esponenziale di media unitaria: il grafico della funzione di rischio cumulato di

	Coef	SE(Coef)	Z	$p-$value	
Peso	0.07	0.03	2.13	0.03	*
Peso^2	-10^{-4}	0.00	-1.87	0.06	.
Allattamento Totale	-1.43	0.60	-2.38	0.02	*
Numero.Bassi	-0.12	0.06	-1.86	0.06	.
Numero.Alti	0.41	0.30	1.38	0.17	
Numero.Picchi	0.05	0.03	1.73	0.08	.
Numero.Bassi^2	10^{-3}	0.00	2.19	0.03	*
Supplementi	1.38	0.76	1.83	0.07	.
Primo Alto	1.17	0.62	1.89	0.06	.
Primo Picco	2.94	0.35	8.35	0.00	***
Allattamento:N°.Alti	0.34	0.24	1.44	0.15	
Allattamento:N°.Picchi	0.32	0.13	2.39	0.02	*
N°.Alti:Supplementi	-0.59	0.26	-2.25	0.02	*

	$IC_{0.95}$ inferiore	Coef	$IC_{0.95}$ superiore
Peso	0.01	0.07	0.13
Peso^2	$-3 \cdot 10^{-4}$	-10^{-4}	$-2 \cdot 10^{-6}$
Allattamento Totale	-2.57	-1.43	-0.22
Numero.Bassi	-0.24	-0.12	$-2 \cdot 10^{-3}$
Numero.Alti	-0.20	0.41	0.94
Numero.Picchi	10^{-3}	0.05	0.093
Numero.Bassi^2	$-2 \cdot 10^{-4}$	10^{-3}	$-3 \cdot 10^{-3}$
Supplementi	-0.17	1.38	2.76
Primo Alto	0.14	1.17	2.50
Primo Picco	2.03	2.94	3.32
Allattamento:N°.Alti	-0.11	0.34	0.82
Allattamento:N°.Picchi	0.02	0.32	0.55
N°.Alti:Supplementi	-1.07	-0.59	-0.06

Tabella 4.1: Coefficienti associati alle variabili esplicative del Modello di Cox con relativi errori standard, significatività ed intervalli di confidenza

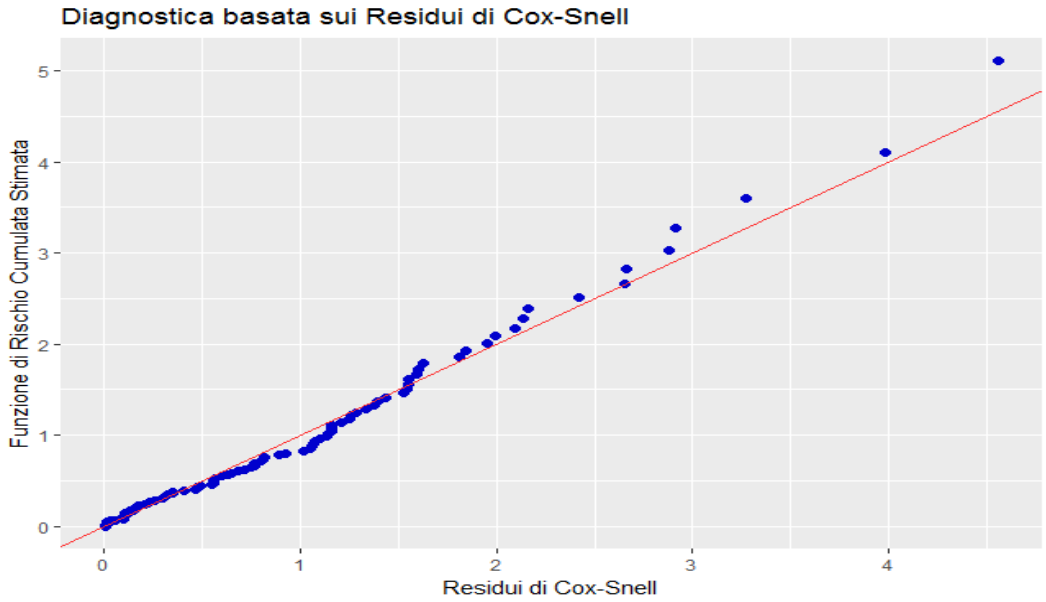


Figura 4.1: Grafico dei residui di Cox-Snell e della stima della loro funzione di rischio cumulata.

questi residui dovrebbe quindi oscillare il più vicino possibile attorno alla bisettrice del primo quadrante.

Nel grafico 4.2 mettiamo a confronto i predittori lineari degli episodi delle diverse unità statistiche coi relativi residui martingala, come proposto in Therneau *et al.* (1990). Pur non essendo un grafico eccessivamente soddisfacente, non notiamo delle irregolarità così pesanti da farci dubitare della correttezza del nostro modello.

Per verificare che l'assunzione di proporzionalità dei rischi sia rispettata, utilizziamo il test proposto da Grambsch e Therneau (1994). Senza addentrarci nei dettagli, in questo test vengono utilizzati i residui di Shoenfeld relativi ad ogni variabile esplicativa per stimare i relativi coefficienti sotto forma di funzioni dipendenti dal tempo $\hat{\beta}_i(t)$ utilizzando dei metodi di lisciamento. È possibile effettuare un test statistico per verificare l'ipotesi che questa funzione sia costante nel tempo. È inoltre possibile ottenere una statistica test relativa al fatto che l'assunzione di rischi proporzionali sia verificata globalmente per il modello a rischi proporzionali adattato. Nella tabella 4.2 sono riportati i valori di queste statistiche test, coi rispettivi p-value. È inoltre riportato il p-value della statistica test che verifica l'assunzione globale di proporzionalità del modello. Tutti i p-value risultano superiori alla soglia di 0.10, non abbiamo quindi elementi per rifiutare l'ipotesi di proporzionalità dei rischi. Nella figura 4.3 sono mostrati i grafici dei residui di Schoenfeld per ogni variabile, con le stime dei differenti $\hat{\beta}(t)$. Come confermatoci dal

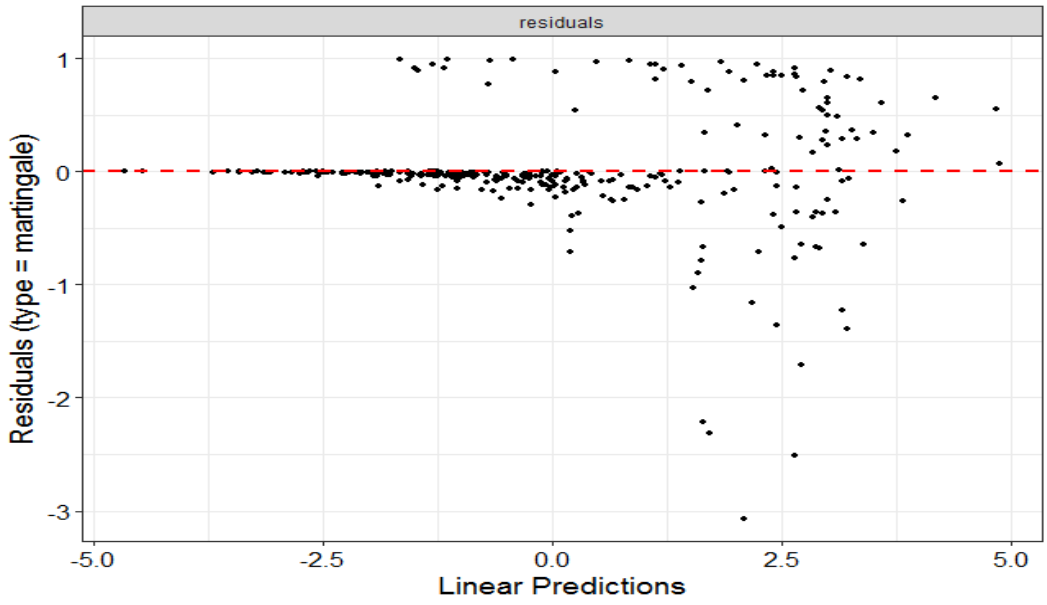


Figura 4.2: Grafico dei predittori lineari contro i residui martingala.

test, non notiamo violazioni dell'assunto di proporzionalità.

4.4.2 Interpretazione dei risultati

I coefficienti ottenuti dal modello a rischi proporzionali rispondono al problema di valutare quale effetto abbiano le variabili esplicative sul rischio di ritorno della fertilità dopo il parto. Notiamo come il peso abbia un effetto sia lineare che quadratico: al crescere del peso aumenta il rischio di sperimentare la fine della fase di amenorrea post parto; l'effetto quadratico invece mitiga leggermente questo effetto.

Un elevato numero di bassi settimanali riduce il rischio di ritorno della fertilità, mentre al crescere del numero di alti e di picchi settimanali questa quantità aumenta. Questo risultato è perfettamente sensato: al crescere della prima variabile si osserva una bassa attività ormonale; l'aumento delle altre due variabili, invece, implica un aumento di questa attività ormonale, evento tipicamente associato al ritorno della fertilità.

Il numero di bassi entra nel modello anche come effetto quadratico: il valore positivo del suo coefficiente mitiga leggermente l'effetto negativo causato dal termine lineare. Il fatto di avere sperimentato il primo alto causa un aumento del rischio del ritorno alla fertilità; l'effetto positivo del picco ormonale sul rischio risulta ancora più elevato, come è sensato aspettarci, dato che questo valore indica l'elevata presenza di ormone luteinizzante, tipicamente associata al momento dell'ovulazione.

	ρ	χ^2	$p-$value
Peso	-0.05	0.16	0.69
Peso^2	0.05	0.19	0.66
Allattamento Totale	-0.00	0.00	0.98
Numero.Bassi	-0.00	0.00	0.98
Numero.Alti	-0.02	0.06	0.81
Numero.Picchi	0.08	0.40	0.53
Numero.Bassi^2	0.01	0.01	0.92
Supplementi	0.01	0.02	0.90
Primo Alto	-0.06	0.34	0.56
Primo Picco	-0.03	0.11	0.75
Allattamento:N°.Alti	-0.01	0.02	0.90
Allattamento:N°.Picchi	0.12	1.52	0.22
N°.Alti:Supplementi	0.02	0.04	0.84
GLOBAL		4.03	0.99

Tabella 4.2: Tabella relativa al test per l'assunzione di proporzionalità dei rischi.

Le variabili che non in questo studio non sono risultate influenti sul rischio del ritorno alla fertilità della donna sono l'età, l'altezza, l'indice di massa corporea (codificato come variabile statica), il numero di gravidanze sperimentate ed il numero di figli attualmente in vita.

Passiamo ora all'esame delle due variabili di maggior interesse del dataset, nonché alle più delicate da trattare. Il fatto di avere introdotto i supplementi nella dieta del bambino aumenta sensibilmente il rischio di ritorno alla fertilità: questo effetto viene tuttavia mitigato dall'interazione con la media degli alti settimanali. Nel caso in cui ci sia una media settimanale superiore a circa 3 alti, caratteristica presente circa nel 38% del campione, l'effetto moltiplicativo positivo che questa variabile induce sul rischio viene annullato ed, anzi, sembrerebbe causare un ritorno alla fertilità più lento. Allo stesso modo, anche il fatto di aver interrotto l'allattamento esclusivo al seno porta alla crescita del rischio di ritorno della fertilità: questo effetto viene mitigato dall'interazione con la media del numero di alti e di picchi settimanali. In particolare, nel 43% dei soggetti l'effetto positivo delle interazioni supera l'effetto negativo associato all'allattamento esclusivo, causando di fatto un effetto totale positivo sul rischio.

Le interazioni sembrerebbero cogliere l'effetto di selezione delle unità causato dalla dicotomizzazione della variabile relativa all'allattamento

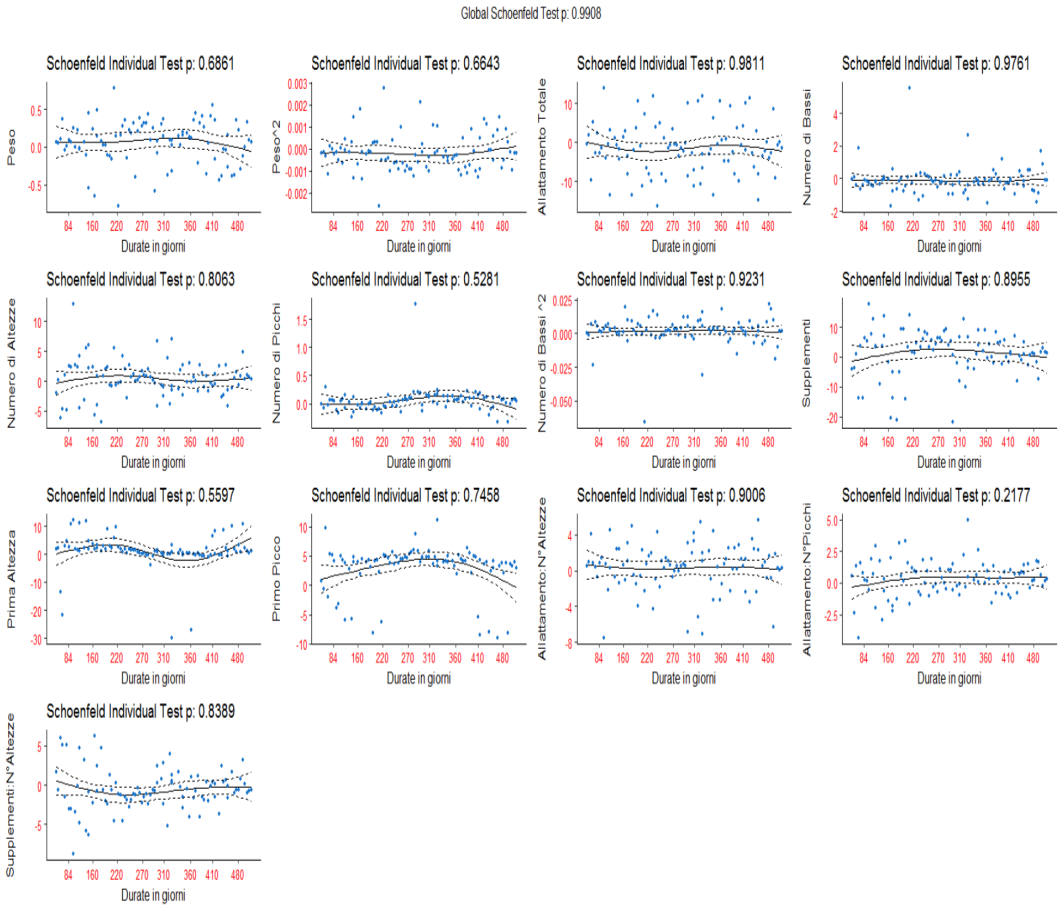


Figura 4.3: Grafici relativi al test per l'assunzione di proporzionalità dei rischi.

totale. Infatti, la distribuzione delle durate relative alle unità con effetto sul rischio positivo indotto dall'allattamento risulta bimodale, mentre la distribuzione delle durate relative alle unità con effetto negativo è unimodale. Possiamo interpretare la prima moda di unità con effetto sul rischio positivo come il gruppo delle unità che non hanno avuto la possibilità di interrompere l'allattamento esclusivo al seno a causa del fatto di aver sperimentato il ritorno alla fertilità troppo precocemente. Nella figura 4.4 sono riportati i relativi istogrammi.

Il fatto che queste interazioni risultino statisticamente significative non è immediatamente intuitivo. Al fine di comprendere se questi particolari effetti evidenziati dal nostro modello siano dovuti ad una codifica non ottimale delle variabili o a fattori del tutto casuali, potrebbe essere sensato in un futuro studio aumentare il numero di unità statistiche partecipanti all'analisi, e cercare di raccogliere le relative variabili esplicative in maniera più dettagliata. Alla luce di queste nuove osservazioni sarà più semplice stabilire se i risultati ottenuti da questo studio possano

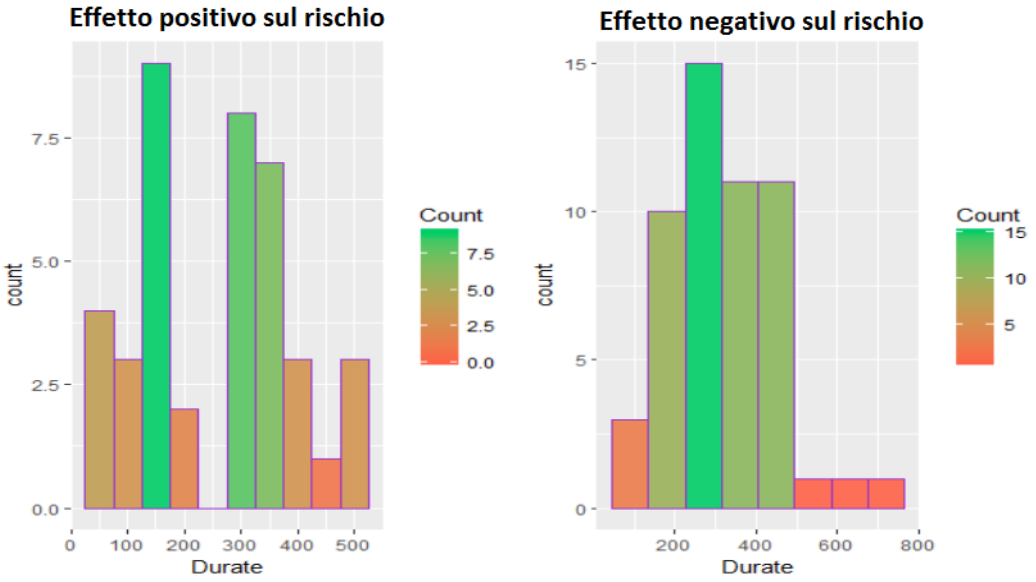


Figura 4.4: Istogrammi relativi alle durate dei due sottogruppi individuati dall'effetto complessivo dell'allattamento sul rischio.

essere considerati soddisfacenti ed essere riconosciuti come validi dalla comunità scientifica, oppure essere semplicemente ignorati.

4.5 Modelli alternativi

Il modello a rischi proporzionali si adatta molto bene ai dati: abbiamo comunque provato ad utilizzare altri tipi di modelli nel corso delle analisi. Uno dei primi modelli alternativi che abbiamo provato ad adattare ai dati è stato un modello di *Survival Random Forest* (Ishwaran *et al.*, 2008), una particolare tipologia di Foresta Casuale formulata esplicitamente per trattare dati di durata. Questo modello tuttavia non si è rivelato utile per il nostro studio, infatti non è in grado di gestire variabili dipendenti dal tempo, caratteristica essenziale per analizzare il dataset in questione.

Abbiamo provando comunque ad utilizzarlo adattando al dataset senza le tre variabili dipendenti dal tempo sia un modello a rischi proporzionali che una Survival Random Forest. Per valutare quale dei due modelli si adattasse meglio ai dati, abbiamo confrontato i valori dei *C-index* (Triantaphyllou, 2000) ottenuti per i due modelli: il modello di Cox ci ha fornito un valore di C-index più alto, risulta quindi preferibile al modello di Foresta Causale. Probabilmente, data la non elevata numerosità campionaria, un modello semiparametrico in cui le assunzioni

siano rispettate riesce ad essere più efficiente rispetto ad un modello completamente non parametrico.

Un altra tipologia di modelli alternativi che abbiamo provato ad adattare ai dati è quella dei modelli a rischi additivi.

Il primo modello di questo tipo che abbiamo provato ad utilizzare è il modello non parametrico a rischi additivi di Aalen (Aalen, 1980). In questa particolare formulazione, la funzione di rischio associata ad ogni unità è modellata nella seguente maniera:

$$\lambda_i(t|x) = \alpha_0(t) + \alpha(t)^T \mathbf{x}_i(t), \tag{4.6}$$

con $\lambda_i(\cdot)$ funzione di rischio per l'i-esima osservazione, $\mathbf{x}_i$ vettore delle relative covariate e $\alpha(t)$ vettore di funzioni dipendenti dal tempo di dimensione $p \times 1$. $\alpha_0(t)$ è la funzione dipendente dal tempo relativa all'intercetta: tipicamente in questo modello si aggiunge alla matrice delle covariate un'ulteriore variabile esplicativa a valori unitari al fine di stimare l'intercetta. Nonostante questo modello sia ideale per trattare variabili dipendenti dal tempo, non si è rivelato utile nella nostra analisi: avendo a disposizione solamente 93 unità statistiche, ottenere una stima non parametrica per tutte e 13 le variabili regressive considerate si è rilevato infattibile.

4.5.1 Modello semiparametrico a rischi additivi

Un modello additivo più parsimonioso è il modello semiparametrico a rischi additivi introdotto da McKeague e Sasieni (1994). La formulazione della funzione di rischio è simile al modello di Aalen, con la differenza che alcuni coefficienti sono trattati come dipendenti dal tempo, mentre altri sono considerati fissi nel tempo. Nello specifico:

$$\lambda_i(t|x, z) = \alpha_0(t) + \alpha(t)^T \mathbf{x}_i(t) + \beta^T \mathbf{z}_i(t), \tag{4.7}$$

con $\alpha_0(t)$ intercetta dipendente dal tempo, $\alpha(t)$ vettore degli effetti dipendenti dal tempo, β vettore dei parametri costanti nel tempo, $\mathbf{x}_i$ vettore delle covariate il cui effetto cambia nel tempo, $\mathbf{z}_i$ vettore delle variabili esplicative ad effetti costanti. Nella pratica, invece di stimare il vettore delle funzioni $\alpha(t)$, si stima il vettore degli effetti cumulati $A(t) = \int_0^t \alpha(\tau)d\tau$. Per ogni variabile a cui associamo un effetto dipendente dal tempo è possibile effettuare due particolari test statistici per verificare le ipotesi che l'effetto complessivo sia nullo oppure che sia costante nel tempo (Martinussen e Scheike, 2007).

Il procedimento descritto dagli autori per individuare quali variabili siano costanti nel tempo consiste nell'adattare inizialmente il modello includendo tutte le variabili esplicative nel gruppo di quelle con effetto dipendente dal tempo, quindi di utilizzare per ogni variabile il grafico di $A_i(t)$ al variare del tempo ed il test sopracitato per valutare quali variabili abbiano un effetto costante nel tempo. Una volta individuate queste variabili, il modello viene riadattato includendo le covariate identificate dal test precedente nel gruppo delle variabili a coefficienti costanti. Il procedimento da noi eseguito è molto simile a quello suggerito: non potendo però adattare un modello non parametrico completo su cui eseguire il test di dipendenza dal tempo a causa della ridotta numerosità campionaria, abbiamo inserito una variabile alla volta. Ogni nuova variabile inserita nel modello è stata inizialmente inclusa come dipendente dal tempo, per poi escluderla dai modelli nel caso in cui l'ipotesi del test di nullità degli effetti venisse accettata, oppure inserirla nel modello successivo come variabile a coefficienti costanti nel tempo nei casi individuati dal test per gli effetti costanti nel tempo.

Nel nostro caso, tutte le variabili inserite nel modello sono risultate avere effetti costanti nel tempo, a conferma della bontà del modello a rischi proporzionali da noi formulato in precedenza. Il modello semiparametrico a rischi additivi ha individuato come significative solamente le variabili "Peso","Numero di alti", "Numero di Picchi", "Primo Picco Registrato": con l'eccezione della variabile "Primo Alto", queste sono esattamente le stesse variabili che la procedura di *stepwise backward* eseguita sul modello di Cox senza termini quadratici o interazioni suggerisce di includere nel modello. Come nel modello a rischi proporzionali, anche questi coefficienti influiscono positivamente sul rischio di sperimentare il ritorno alla fertilità.

Capitolo 5

Ritorno della fertilità senza picchi ormonali: analisi statistiche

In questo capitolo verranno presentate le analisi statistiche effettuate per rispondere alla seconda domanda posta nella sezione 2.4: è di interesse comprendere quali siano i fattori che influenzano il fatto di non osservare picchi di ormone luteinizzante durante la fase di amenorrea post-parto, e quindi di non riuscire ad ottenere una stima del giorno di ovulazione sulla base dell'esito del monitoraggio elettronico.

5.1 Modelli a rischi competitivi

Il fatto che il ritorno alla fertilità possa avvenire senza che il dispositivo *EHFM* abbia rilevato la presenza di almeno un picco ormonale durante tutta la durata della fase di amenorrea post parto può essere un fattore problematico per l'efficacia del protocollo *"Marquette Method"*. Per questa ragione, i ricercatori della Marquette University sono interessati a comprendere quali siano i fattori che influiscono sul fatto che ci sia almeno un picco di ormone LH durante la fase di assenza della fertilità.

Prima di effettuare le analisi statistiche, è necessario comprendere la tipologia di problema che dobbiamo trattare. Al contrario della domanda affrontata nel Capitolo 4, in questo caso per ogni unità statistica possono avvenire due eventi di interesse distinti: la segnalazione di un picco ormonale, utilizzato per ricavare la stima della data dell'ovulazione, e l'evento relativo al ritorno alla fertilità.

Un possibile primo approccio per affrontare questo quesito è di trattarlo come un problema di dati di durata a rischi competitivi. Questa particolare tipologia di analisi di sopravvivenza è stata molto trattata in letteratura, data la frequente necessità di affrontare problemi in cui possano accadere alla stessa unità statistica più eventi d'interesse. Si vedano, ad esempio, Gail (1975), Xu (1999), Hougaard (2012).

Un problema a rischi competitivi è formulato nella seguente maniera: per ogni unità statistica non abbiamo più un unico evento di interesse, che può effettivamente avvenire oppure non essere osservato andando così a creare una durata censurata, ma diversi eventi finali possibili. Un soggetto esce dall'insieme delle unità a rischio nel caso in cui sperimenti uno di questi eventi oppure sia censurato: in ogni caso, una volta uscito non sarà più possibile ottenere informazioni sul quella particolare unità statistica. Come vedremo, in certe situazioni questo fatto può essere un limite.

5.1.1 Creazione del dataset

Definiamo nel nostro studio i due eventi competitivi "Osservazione del primo picco ormonale" e "Ritorno della Fertilità dopo il parto": abbiamo infatti a nostra disposizione le durate relative a questi due eventi per buona parte delle donne incluse nello studio. La variabile risposta per ogni soggetto è definita come la durata minima tra il tempo di ritorno alla fertilità dopo il parto ed il tempo trascorso tra la nascita del bambino e la prima rilevazione di un picco ormonale. Abbiamo scelto di utilizzare come evento competitivo la variabile "Primo Picco Registrato" al posto della variabile "EDO". Questa scelta è stata motivata dal fatto che una volta osservato il primo picco non sia più possibile osservare il ritorno alla fertilità senza ottenere una stima del giorno di ovulazione. La variabile "EDO" è stata definita come l'ultimo picco ormonale rilevato prima della prima mestruazione fertile: se decidessimo di utilizzare questa variabile come evento competitivo, non potremmo affermare che questo sia avvenuto senza aver prima osservato l'evento relativo al ritorno della fertilità. Nel dataset originale abbiamo una variabile dicotomica che ci dice se nel Ciclo 0 sia stata registrata una stima del giorno di ovulazione oppure no. Dopo la pulizia del dataset e gli aggiustamenti descritti nella sezione 3.1, questa variabile presentava circa il 20% di dati mancanti. Si è scelto di escludere questi soggetti dalle analisi statistiche successive, in quanto per queste unità statistiche non abbiamo l'informazione sul fatto che sia effettivamente avvenuto un picco di ormone LH prima

del ritorno alla fertilità. Sarebbe stato possibile ottenere questi valori utilizzando i dati imputati durante l'analisi del problema precedente, ma abbiamo fatto la scelta conservativa di non includere nello studio variabili risposta imputate, nonostante la perdita di 19 soggetti sui 93 totali. Dopo questa operazione, infatti, la numerosità del campione si è ridotta a 74 unità. Come nel problema precedente, anche in questo caso è presente una sola osservazione con durata censurata, tutte le altre unità sperimentano uno dei due eventi di interesse.

Come variabili esplicative usiamo le variabili contenute nel dataset impiegato per l'analisi del problema precedente, anche se non possiamo più utilizzare le tre variabili relative al numero di bassi, alti e picchi registrati durante il Ciclo 0 in quanto non abbiamo informazione su quanti di questi siano stati registrati prima dell'osservazione del primo picco ormonale e quali successivamente. Le altre variabili sono rimaste inalterate, con l'eccezione della variabile relativa alla data del primo picco che è diventata parte della variabile risposta. Le variabili esplicative utilizzate per rispondere a questo problema sono quindi "Età", "Altezza", "Peso", "BMI", "Numero di Gravidanze", "Figli Attualmente in Vita", "Allattamento Totale", "Supplementi", "Primo Alto".

I valori mancanti delle variabili esplicative sono stati ricavati prendendo i valori ottenuti dall'algoritmo *missForest* durante l'imputazione del dataset utilizzato durante le analisi svolte nel Capitolo 4.

Delle 74 unità statistiche a nostra disposizione, 60 sperimentano l'evento "Primo Picco Ormonale", mentre solamente 13 soggetti hanno sperimentato il ritorno alla fertilità senza aver ottenuto un EDO. Il fatto di avere una numerosità così bassa per uno dei due eventi potrebbe essere un problema. Vedremo in ogni caso nella prossima sezione cosa ci diranno i modelli adattati.

5.1.2 Modelli statistici a rischi competitivi: esposizione ed adattamento

L'approccio classico per l'analisi di un problema a rischi competitivi, denominato *"a cause specifiche"* (Prentice *et al.*, 1978), prevede di adattare per ognuno dei K eventi di interesse K modelli differenti, al fine di modellare le relative funzioni di rischio. È stato infatti dimostrato come nella funzione di verosimiglianza globale il fattore relativo al k-esimo evento sia esattamente pari alla funzione di verosimiglianza dell'evento stesso che si otterrebbe nel caso in cui le realizzazioni di tutti gli altri eventi competitivi fossero censurati. È quindi possibile utilizzare i clas-

sici modelli di analisi di dati di durata per modellare le diverse funzioni di rischio a causa specifica semplicemente adattando K modelli e trattando come casi censurati tutti gli eventi differenti dal k-esimo evento di interesse. Anche in questo caso, un modello molto popolare è il modello a rischi proporzionali di Cox: la formulazione della k-esima funzione di rischio a causa specifica per l'i-esima osservazione, con i=1, ... n e k=1, ..., K, può essere espressa nella seguente maniera:

$$\lambda_{ki}(t) = \lambda_{k0}(t)e^{\mathbf{x}_i\boldsymbol{\beta}_k}, \tag{5.1}$$

con $\lambda_{k0}(t)$ funzione di rischio di base a causa specifica per il k-esimo evento e $\boldsymbol{\beta}_k$ vettore di dimensione $p\times 1$ dei relativi coefficienti. Nell'adattare uno di questi k modelli, le durate relative ai casi che sperimentano un evento differente da quello di interesse vengono lasciate inalterate, ma vengono considerati come censurati. Tutte le usuali proprietà del modello di Cox per singoli eventi, i metodi di stima e di verifica di buon adattamento ai dati rimangono inalterati.

Questo approccio, per quanto semplice, risulta molto efficace; presenta tuttavia un limite che nel corso degli anni si è rilevato sempre più preponderante. In molti problemi a rischi competitivi risulta di grande interesse per i ricercatori avere a disposizione uno strumento per stimare la probabilità di sperimentare uno dei k esiti al tempo t (Pepe e Mori, 1993). Questa quantità può essere ottenuta mediante la funzione di incidenza cumulativa (Putter *et al.*, 2007), definita per l'evento k come $p_k(t) = \int_0^t \lambda_k(s)S(s)ds$, con $\lambda_k(\cdot)$ funzione di rischio a causa specifica per l'evento k ed $S(\cdot)$ funzione di sopravvivenza complessiva, definita nella seguente maniera:

$$S(t) = e^{-\sum_{k=1}^{K}\Lambda_k(t)dt}, \tag{5.2}$$

con $\Lambda_k(t)$ funzione di rischio cumulata a causa specifica per il k-esimo evento. Le funzioni di rischio ottenute mediante un approccio a causa specifica possono essere utilizzate per stimare marginalmente queste probabilità, tuttavia le informazioni ottenute marginalmente da queste funzioni possono essere molto diverse dai risultati ottenuti dalla funzione di incidenza cumulata, che considera le probabilità complessive di tutti gli eventi possibili (Gray, 1988). Per ottenere delle stime della funzione di rischio che forniscano gli stessi risultati della funzione di incidenza cumulata, Fine e Gray (1999) hanno proposto un modello a rischi

proporzionali simile al modello di Cox. La funzione di verosimiglianza parziale utilizzata per stimare i parametri del modello si avvale di un insieme di rischio diverso rispetto al classico modello a rischi proporzionali: mentre il contributo delle unità censurate o che sperimentano l'evento d'interesse rimane inalterato, le unità che invece sperimentano eventi differenti da quello considerato nel k-esimo modello restano nell'insieme dei soggetti a rischio anche dopo aver sperimentato l'evento. A queste osservazioni vengono associati dei pesi via via decrescenti, ottenuti mediante una tecnica di probabilità inversa dei pesi delle censure (*"Inverse Probability of Censoring Weighting"*, IPCW) (Robins e Rotnitzky, 1992).

Nelle nostre analisi, abbiamo provato ad utilizzare entrambi gli approcci. Per ottenere il dataset su cui adattare il modello a cause specifiche è stato sufficiente effettuare sul dataset descritto nella sezione 5.1.1 la procedura di suddivisione delle osservazioni in sotto-episodi descritta nella sezione 4.2 per trattare le variabili dipendenti dal tempo. Sono stati quindi adattati due modelli a rischi proporzionali differenti ponendo di volta in volta i casi che sperimentano l'evento non di interesse come censurati.

Per adattare il modello di Fine-Gray, invece, è stato necessario creare due dataset distinti. Su ognuno di questi abbiamo eseguito le stesse operazioni effettuate per il modello precedente, aggiungendo per ogni unità che ha sperimentato l'evento alternativo una serie di sotto-episodi censurati successivi al verificarsi del suo evento. A ciascuno di questi sono stati assegnati dei pesi decrescenti, ottenuti mediante la tecnica di IPCW.

Entrambe le tipologie di modelli adattati hanno fornito esiti molto simili, com'era da aspettarsi. Tuttavia, i risultati non si sono rilevati molto informativi. Gli unici due fattori che sono risultati significativi per il fatto di osservare un picco ormonale sono il peso e l'osservazione del primo alto, i quali hanno entrambi un effetto positivo sulla funzione di rischio. Non è stato invece possibile individuare alcun fattore che influenzi il fatto di osservare il ritorno alla fertilità senza aver registrato un picco di ormone luteinizzante. Non siamo sicuri se questo sia un problema di metodo utilizzato, di scarsa numerosità campionaria per le unità che non hanno sperimentato un picco ormonale, o semplicemente perché effettivamente non ci siano fattori che influenzano il rischio di sperimentare questo evento, almeno tra le variabili incluse nel dataset in questione.

Nella sezione successiva formuleremo il problema statistico in ma-

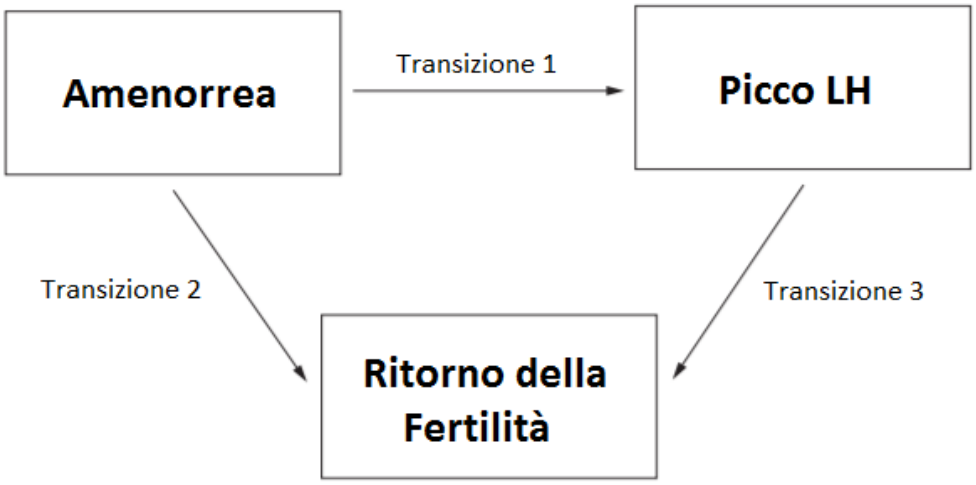


Figura 5.1: Schema delle possibili transizioni di stato del modello Malattia-Morte

niera differente, cercando di ottenere dei risultati più soddisfacenti.

5.2 Modello "Malattia-Morte"

La scarsa numerosità campionaria dei soggetti che sperimentano il ritorno della fertilità senza aver subito picchi ormonali può essere una delle cause del non essere riusciti ad individuare alcun fattore che abbia un certo effetto sul fatto di sperimentare questo particolare evento. Proviamo allora ad includere nel modello l'informazione relativa al periodo compreso tra il primo picco ormonale ed il ritorno alla fertilità, sperando in questo modo di ottenere un modello più completo.

Il modello che utilizzeremo per risolvere questo problema è denominato "Modello di Malattia-Morte", dato che le prime applicazioni in cui è stato formulato riguardavano problemi di questo tipo (Fix e Neyman, 1951; Sverdrup, 1965). Certe volte, è indicato in letteratura anche come modello "Disabilità" (Hougaard, 1999). In questo modello le unità possono trovarsi in tre stati possibili: lo stato iniziale, uno stato intermedio e lo stato assorbente. Le unità si possono muovere nei diversi stati al variare del tempo, seguendo dei percorsi prefissati. L'ingresso in un nuovo stato è definito dal verificarsi di un prefissato evento di interesse. Le unità che si trovano nello stato iniziale possono transitare o nello stadio finale ("Morte"), oppure transitare allo stato intermedio ("Malattia"). Le unità nello stato finale non possono più transitare in alcuno stato: escono in quel momento dallo studio e non è più possibile

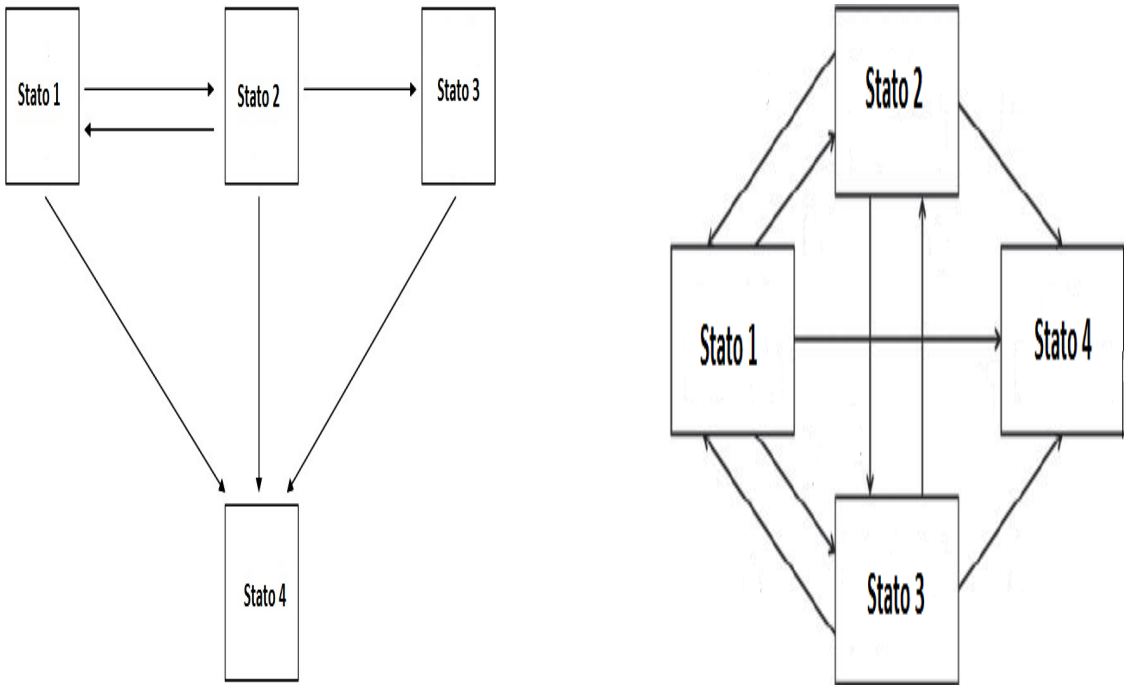


Figura 5.2: Alcuni esempi di possibili grafici di modelli multi-stato

ottenere informazioni su di esse. Per definizione, tutte le unità entrano nello studio nello stato iniziale al tempo 0. Certe volte è permessa anche la transizione dallo stadio intermedio allo stato iniziale: nel nostro caso, questa transizione non è possibile.

Nella figura 5.1 osserviamo una schematizzazione dei possibili stati in cui le unità si possono trovare, insieme alle relative transizioni possibili. Denominiamo la transizione dallo stato iniziale all'osservazione del primo picco ormonale "Transizione 1", la transizione che va direttamente dallo stato iniziale al ritorno della fertilità "Transizione 2", la transizione dalla rilevazione del primo picco alla fine della fase di amenorrea post-parto "Transizione 3".

5.2.1 Modelli multi-stato

Il modello malattia-morte introdotto nella sezione precedente è un caso particolare di una classe di modelli più ampia: i Modelli Multi-Stato. In questa particolare tipologia di modelli, definiamo il processo stocastico $\mathbf{X}(t)$ in funzione del tempo t, con spazio degli stati $S=\{1,\ldots,K\}$ (Meira-Machado *et al.*, 2008). Le singole unità statistiche si muovono al variare del tempo da uno stato all'altro, seguendo un percorso definito dall'insieme delle possibili transizioni da uno stato all'altro. Il cambio di stato è demarcato dal verificarsi di uno dei $J < K$ possibili

eventi di interesse per la transizione corrente. Tutte le unità entrano nello studio in un certo stato iniziale, e vi escono una volta entrati in uno stato assorbente, da cui non possono più uscire. Non è possibile ottenere informazioni per un'unità statistica una volta entrata in uno stato assorbente. A seconda di come è stato definito il modello, possono esistere più stati iniziali e più stati assorbenti. È possibile riassumere tutti gli stati con le possibili transizioni di un modello multi-stato in un grafico: nella figura 5.2 sono riportati alcuni esempi di possibili modelli multi-stato (Hougaard, 1999). Ad ogni possibile transizione è associata una particolare funzione di rischio, al fine di modellare il meccanismo che regola le durate di permanenza nel singolo stato prima di passare a quello successivo. Anche il modello a rischi competitivi può essere visto come un caso particolare di modello multi-stato (Andersen *et al.*, 2002).

Una delle formulazioni più popolari per modellare le K funzioni di rischio relative alle K diverse transizioni è di utilizzare una formulazione a rischi proporzionali: per la k-esima transizione relativa all'i-esima unità, con k=1, ..., K e i=1, ... n è la seguente:

$$\lambda_{ki}(t) = \lambda_{k0}(t)e^{\mathbf{x}_i\boldsymbol{\beta}_k}. \tag{5.3}$$

È possibile stimare le K funzioni di rischio mediante un singolo modello a rischi proporzionati stratificato, utilizzando come variabile di stratificazione l'etichetta della transizione di interesse. Sarà tuttavia necessario effettuare diverse operazioni di preparazione sul dataset di partenza. L'osservazione nel dataset relativa all'i-esima unità statistica viene suddivisa in K sotto-episodi distinti, uno per ogni possibile transizione. In ognuno di essi vengono registrati il tempo di ingresso e di uscita dal k-esimo stato, il fatto che l'evento osservato sia censurato o meno, e le rispettive variabili esplicative. Se un'osservazione non visita il k-esimo stato, viene comunque associato un episodio di durata nulla a questa transizione. Nel caso di variabili dipendenti dal tempo, il singolo sotto-episodio viene a sua volta diviso in un certo numero di sotto-episodi temporali, definiti sulla base delle variazioni dei valori delle variabili dipendenti dal tempo, analogamente alla procedura descritta nella sezione 4.2.

La stima dei parametri avviene mediante la ben nota procedura di massimizzazione della verosimiglianza parziale, tuttavia l'insieme dei soggetti a rischio al tempo t risulta differente dal classico insieme di rischio per il modello di Cox ad evento singolo. Nel caso in cui un'unità si trovi in uno stato k con una sola transizione possibile il suo contributo alla verosimiglianza parziale per il k-esimo stato comincerà nell'istante

temporale in cui è entrato nello stato corrente e termina nel momento in cui vi esce, come nel caso di un modello di Cox per un singolo evento di interesse con episodi separati. Se invece il soggetto si trova in uno stato in cui è possibile transitare in $1 < J < K$ stati differenti, verranno create J unità artificiali tutte facenti parte dell'insieme di rischio. Nel momento in cui si verifica uno degli eventi di interesse, tutte queste unità escono dall'insieme dei soggetti a rischio: il caso relativo all'evento verificatosi viene considerato come un caso completo, tutti gli altri come casi censurati.

Notiamo come, nell'equazione 5.3, il vettore dei parametri sia relativo alla k-esima transizione: è infatti possibile stimare un vettore di coefficienti per ogni transizione mediante il metodo introdotto in Andersen *et al.* (1991). Ogni variabile esplicativa viene suddivisa in K diverse variabili, ognuna di queste fa riferimento ad una ed una sola transizione. Nel caso in cui un'osservazione sia relativa alla k-esima transizione, il valore originale della k-esima variabile esplicativa così formata viene mantenuto, se invece l'episodio in questione è relativo ad una delle $k-1$ transizioni restanti, il suo valore viene posto come nullo. Da notare come, mentre per le variabili numeriche questo metodo non comporti problemi, nel caso di variabili categoriali si possa creare una certa ambiguità. È comunque pratica comune utilizzare lo stesso approccio anche per questo tipo di covariate, si veda ad esempio Lando e Skødeberg (2002) o Putter *et al.* (2006).

Prima di passare all'adattamento dei modelli, è necessario discutere di un aspetto molto importante relativo al modello in questione. Il modello di malattia-morte può essere adattato nella formulazione presentata in precedenza solamente nel caso sia verificata l'assunzione di markovianità (Meira-Machado *et al.*, 2008). Dato il processo $X(t)$ con spazio degli stati $S = \{1, \ldots, K\}$ al tempo t e due istanti temporali s e t, con $0 < s < t$, definiamo H_{s-} come la storia relativa al processo $X(t)$, ossia l'insieme di tutti gli stati visitati dal processo insieme ai tempi di entrata ed uscita. Il modello è di tipo Markoviano se, dati i due stati j, h in S, vale la seguente proprietà:

$$p(X(t) = j | X(s) = h, H_{s-}) = p(X(t) = j | X(s) = h).$$

Nel caso in cui questa proprietà non sia verificata, non è possibile adattare il modello nella maniera spiegata in precedenza. Sarà necessario includere in qualche modo le informazioni relative agli stati visitati in precedenza ed ai relativi intervalli temporali (Andersen e Keiding, 2002).

Nel modello malattia-morte solamente la terza transizione può violare questa assunzione, dato che le altre cominciano al tempo 0 e non possono quindi avere storia pregressa.

5.2.2 Modello multi-stato a rischi proporzionali: dettagli ed interpretazione

Passiamo adesso all'analisi del dataset mediante l'utilizzo del modello multi-stato di tipo malattia-morte. Il fatto di avere tre transizioni differenti nel modello implica l'esistenza di tre distinte funzioni di rischio da modellare. Anche in questo caso decidiamo di utilizzare il modello semiparametrico di Cox, dato che in tutte le analisi precedenti l'assunzione di proporzionalità dei rischi si è sempre rilevata corretta.

Notiamo la forte analogia con la formulazione della funzione di rischio del modello a rischi competitivi trattato nella sezione 5.1. Il dataset su cui è stato adattato il modello è quello utilizzato per il problema a rischi competitivi esposto nella sezione 5.1. Tuttavia, è stato necessario eseguire le diverse operazioni di suddivisione in sotto-episodi descritte nella sezione precedente.

Ogni osservazione è stata suddivisa in K diversi episodi, uno per ogni possibile transizione, dopodiché ognuno di questi è stato suddiviso in ulteriori sotto-episodi sulla base dei valori delle due variabili dipendenti dal tempo "Primo Alto Osservato" ed "Introduzione dei Supplementi", nel caso in cui durante quel particolare sotto-episodio si verificasse almeno uno di questi due eventi.

Come già affermato in precedenza, le analisi statistiche sono state svolte mediante il linguaggio di programmazione *R*. Le principali funzioni utilizzate per effettuare queste operazioni sul dataset sono state estratte dal pacchetto `mstate` (de Wreede *et al.*, 2011; Putter, 2016). Non abbiamo trovato in letteratura alcun caso di modello malattia-morte adattato utilizzando variabili dipendenti dal tempo, e nel pacchetto utilizzato non era prevista questa possibilità nelle diverse funzioni per gestire i dataset: creare delle funzioni per combinare la suddivisione in sotto-episodi dovuta alle diverse transizioni con la suddivisione dovuta alle variabili dipendenti dal tempo in modo che solamente i corretti sotto-episodi fossero nell'insieme dei soggetti a rischio al tempo t si è rivelato piuttosto complesso, ma alla fine le abbiamo scritte senza problemi.

Scegliendo la formulazione a rischi proporzionali per modellare le funzioni di rischio delle diverse transizioni è possibile utilizzare tutti gli

	Coef	SE(Coef)	Z	p-value	
Peso.1 ^{2}	-10^{-5}	$> 10^{-5}$	2.16	0.03	*
Peso.3	0.17	0.04	4.10	0.00	***
Peso.3 ^{2}	-10^{-4}	$> 10^{-4}$	-4.06	0.00	***
BMI.3	0.19	0.09	2.20	0.03	*
Numero.Gravidanze.3	-0.19	0.07	-2.66	0.01	*
Allattamento.2	1.07	0.65	1.65	0.10	.
Primo.Alto.1	3.61	0.64	5.62	0.00	***

Tabella 5.1: Coefficienti e relativi errori standard del modello di malattia-morte senza tener conto della markovianità.

strumenti relativi al modello di Cox per eventi singoli, come ad esempio i diversi grafici diagnostici basati sui residui o il test per la verifica degli assunti di proporzionalità, come mostrato in Putter (2016).

Inizialmente, il modello è stato adattato senza domandarsi se la proprietà di markovianità fosse verificata o meno. Nella tabella 5.1 sono riportati i valori dei coefficienti, con relativi errori standard e p-value. Le variabili sono state selezionate seguento il principio di minimizzazione del criterio di Akaike. Nella tabella, il numero che segue il nome delle variabili si riferisce alla transizione a cui quel particolare coefficiente fa riferimento. La variabile rispetto alla quale il modello è stato stratificato è la variabile relativa alle diverse transizioni. Notiamo come i risultati del modello siano in realtà piuttosto simili ai risultati ottenuti per il modello a rischi competitivi: le variabili che sono risultate influenti sul fatto di sperimentare un picco ormonale sono il primo alto individuato ed il peso, che questa volta entra con un effetto quadratico. Non siamo riusciti ad individuare fattori che influenzino la seconda transizione, con l'eccezione dell'effetto dell'allattamento esclusivo. Molto probabilmente, questa significatività è dovuta semplicemente all'effetto di selezione dei soggetti, dato che il fatto di aumentare il rischio sembra poco plausibile alla luce dei precedenti studi sull'argomento. Per brevità non vengono riportati i grafici relativi ai residui di Cox-Snell, ai residui di martingala e del test di proporzionalità, che risultano in generale soddisfacenti.

Per quanto questo modello sembri buono nel complesso, non può essere utilizzato data la violazione dell'assunzione di markovianità: seguendo il procedimento descritto ad esempio in Meira-Machado *et al.* (2008), proviamo ad inserire nel modello a rischi proporzionali una variabile relativa alla durata del tempo trascorso nello stato precedente. Naturalmente, è possibile fare questa operazione solamente per la Transi-

	Coef	SE(Coef)	Z	p-value	
Durata.Precedente	10^{-3}	0.00	1.82	0.07	.
Peso.1	0.01	0.00	2.13	0.03	*
Peso.3	0.15	0.04	3.47	0.00	***
BMI.3	0.17	0.09	1.91	0.06	.
Numero.Gravidanze.3	-0.18	0.07	-2.46	0.01	*
Allattamento.2	1.07	0.65	1.65	0.10	.
Primo.Alto.1	3.60	0.64	5.60	0.00	***
Peso.3^2	-10^{-4}	0.00	-3.49	0.00	***

Tabella 5.2: Coefficienti e relativi errori standard del modello di malattia-morte corretto per la markovianità.

zione 3, tutte le altre durate precedenti saranno poste pari a 0. Notiamo un effetto positivo e statisticamente significativo ($p = 0.029$) del tempo trascorso nel primo stato sul rischio di sperimentare il ritorno della fertilità durante la terza transizione. L'assunzione di markovianità risulta dunque violata: sarà necessario incorporare nel nostro modello anche la storia passata relativa alla Transizione 3.

	Rho	Chi Quadro	p-value	
Durata.Precedente	-0.15	4.62	0.03	*
Peso.1	0.09	1.20	0.27	
Peso.3	-0.02	0.08	0.78	
BMI.3	-0.06	0.43	0.51	
Numero.Gravidanze.3	0.10	1.12	0.29	
Allattamento.2	-0.07	0.55	0.46	
Primo.Alto.1	-0.04	0.15	0.70	
Peso.3^2	0.05	0.32	0.57	
GLOBAL		9.19	0.33	

Tabella 5.3: Tabella relativa ai test di proporzionalità del modello corretto per l'assunzione di markovianità

Abbiamo quindi modificato il modello precedente incorporando anche la variabile relativa al tempo passato nello stato precedente. Nella tabella 5.2 sono riportati i coefficienti ottenuti da questo modello, con relativi errori standard e p-values. I risultati ottenuti sono molto simili al modello precedente, con l'unica differenza che l'effetto del peso sul rischio della Transizione 1 non entra più in forma quadratica, ma lineare. Abbiamo notato un leggero miglioramento nell'AIC, tuttavia, i grafici diagnostici risultano peggiorati.

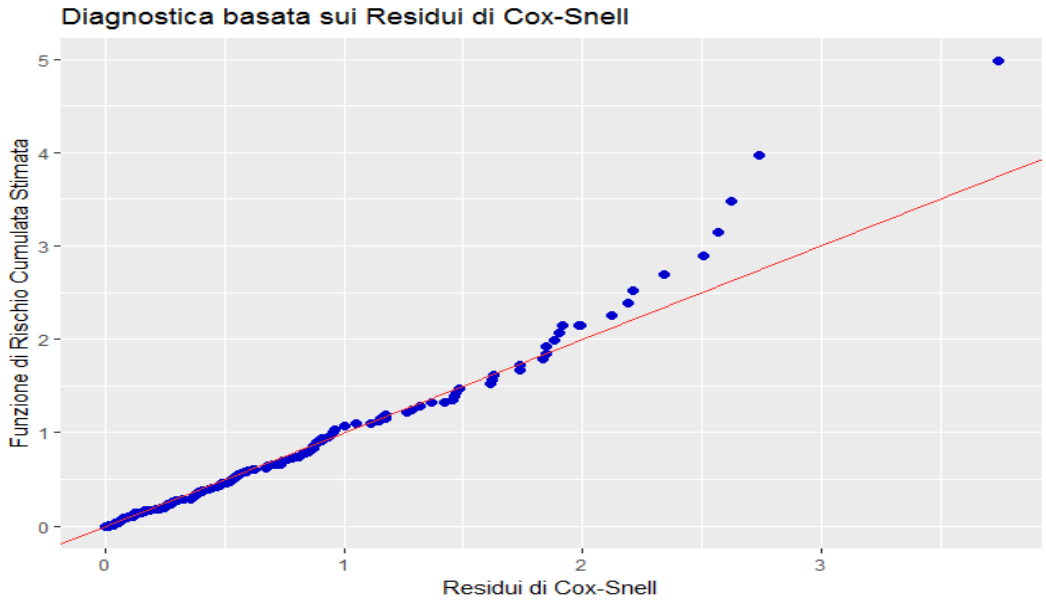


Figura 5.3: Stima della funzione di rischio cumulata per i residui di Cox-Snell per il modello che include le durate passate

Nella figura 5.3 abbiamo riportato il grafico della funzione di rischio cumulata dei residui di Cox-Snell. Notiamo come il grafico si adatti bene fino all'ultimo periodo, in cui il rischio comincia ad essere sovrastimato. Questo andamento si nota anche nel test di proporzionalità: nella tabella 5.3 sono riportati i valori del test, coi rispettivi p-value. Nonostante l'assunzione di rischi proporzionali sia globalmente rispettata, notiamo come il test rifiuti questa ipotesi per la variabile relativa alla durata trascorsa nello stato precedente.

Sembra evidente come l'effetto della variabile relativa alle durate precedenti non sia costante nel tempo. A giudicare dal grafico riportato in figura 5.4, sembra che dopo una certa durata l'influenza del tempo passato diminuisca. Questo effetto è colto anche dal grafico dei residui di Cox-Snell. Il modello utilizzato non è in grado di cogliere l'andamento dipendente dal tempo dell'effetto associato alla suddetta variabile, sarà necessario utilizzare una formulazione per la funzione di rischio che possa modellare coefficienti variabili nel tempo.

5.2.3 Modello multi-stato a rischi additivi

Al fine di cogliere l'effetto dipendente dal tempo per la durata della transizione precedente, modelliamo le funzioni di rischio relative alle diverse transizioni utilizzando il modello semiparametrico a rischi additivi esposto nella sezione 4.5. Il dataset utilizzato rimane invariato.

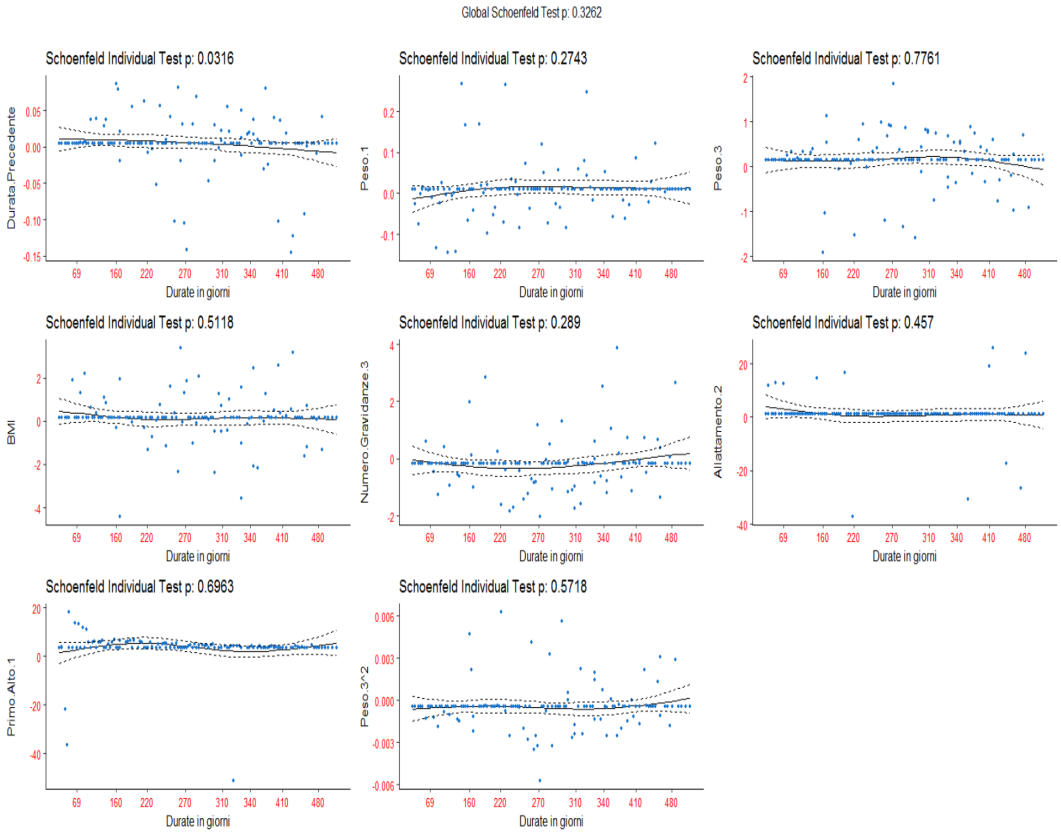


Figura 5.4: Grafici relativi al test per l'assunzione di proporzionalità dei rischi del modello corretto per l'assunzione di markovianità

Nel modello additivo il procedimento di stratificazione è estremamente semplice: è sufficiente trattare la variabile rispetto alla quale si stratifica come una variabile categoriale aggiungendo K variabili dicotomiche per ogni classe presente nella variabile di stratificazione.

La funzione di rischio per l'i-esima unità relativa alla k-esima transizione, con i=1, ... n e k=1, ..., K è espressa nella seguente formulazione:

$$\lambda_{ki}(t) = \alpha_0(t) + \sum_{j=1}^{K} \alpha_{j0}(t) I(j = k) + \alpha_k(t)^T \mathbf{x}_{ik}(t) + \beta_k^T \mathbf{z}_{ik}(t),$$

con $\alpha_0(t)$ intercetta dipendente dal tempo, $\alpha_{j0}(t)$, j=1, ..., K funzione dipendente dal tempo relativa alla j-esima transizione, $\alpha_k(t)$ vettore degli effetti dipendenti dal tempo relativi alla k-esima transizione, β_k vettore dei parametri costanti nel tempo per la k-esima transizione, $\mathbf{z}_{ik}(t)$ e $\mathbf{x}_{ik}(t)$ vettori delle covariate ad effetti rispettivamente costanti e dipendenti dal tempo per la k-esima transizione. I risultati ottenuti

	Coef	SE(Coef)	Z	p-value	
BMI.1	$1.88 \cdot 10^{-4}$	$0.75 \cdot 10^{-4}$	2.51	0.012	*
BMI.2	$0.34 \cdot 10^{-4}$	$0.13 \cdot 10^{-4}$	2.67	0.008	**
BMI.3	$0.63 \cdot 10^{-3}$	$0.20 \cdot 10^{-3}$	3.08	0.002	**
Altezza.1	$-0.55 \cdot 10^{-4}$	$0.28 \cdot 10^{-4}$	-1.98	0.048	*
Primo.Alto.1	$0.89 \cdot 10^{-2}$	$0.11 \cdot 10^{-2}$	7.50	0.000	***
Primo.Alto.2	$0.13 \cdot 10^{-2}$	$0.05 \cdot 10^{-2}$	2.28	0.022	*

Tabella 5.4: Tabella dei coefficienti fissi del modello semiparametrico a rischi additivi, con relativo errore standard e p-value.

da questo modello risultano più interessanti rispetto a quelli ottenuti dal modello a rischi proporzionali: abbiamo infatti individuato delle variabili significative per la Transizione 2. Le variabili individuate come significative e ad effetti dipendenti dal tempo sono l'allattamento relativo alla Transizione 3 e la durata del periodo trascorso nella transizione precedente. Le variabili con effetto costante significative, invece, sono l'indice di massa corporea per tutte e tre le transizioni, il primo alto registrato per le prime due transizioni e l'altezza della donna per la prima transizione. Nella tabella 5.4 sono riportati i valori dei coefficienti fissi con relativi errori standard e p-value, mentre nel grafico 5.5 riportiamo i grafici degli effetti cumulati al variare del tempo. Il test di nullità degli effetti dipendenti dal tempo accetta l'ipotesi nulla per lo strato relativo alla Transizione 2: la prima e la seconda transizione condividono quindi la stessa funzione di rischio di base, mentre quello relativo alla Transizione 3 risulta differente.

Notiamo come il coefficiente relativo all'Indice di Massa Corporea per la prima transizione sia molto più elevato rispetto a quella associato alla seconda transizione, stesso discorso per i coefficienti associati al primo alto. Possiamo interpretare questo fatto affermando che al crescere del BMI aumenta il rischio di sperimentare un picco di ormone luteinizzante prima del ritorno della fertilità. Al crescere dell'altezza, invece, il rischio di sperimentare questo evento diminuisce: il risultato è coerente, considerando la relazione che lega altezza ed indice di massa corporea. Allo stesso modo, il fatto di aver rilevato il primo alto ormonale rende più probabile la transizione verso l'evento "Primo Picco Ormonale". I grafici dei coefficienti cumulati, invece, forniscono un altro tipo di informazione: durante la terza transizione l'allattamento esclusivo al seno sembrerebbe causare un aumento del rischio, effetto probabilmente dovuto al meccanismo di selezione spiegato in precedenza. Notiamo invece come, al

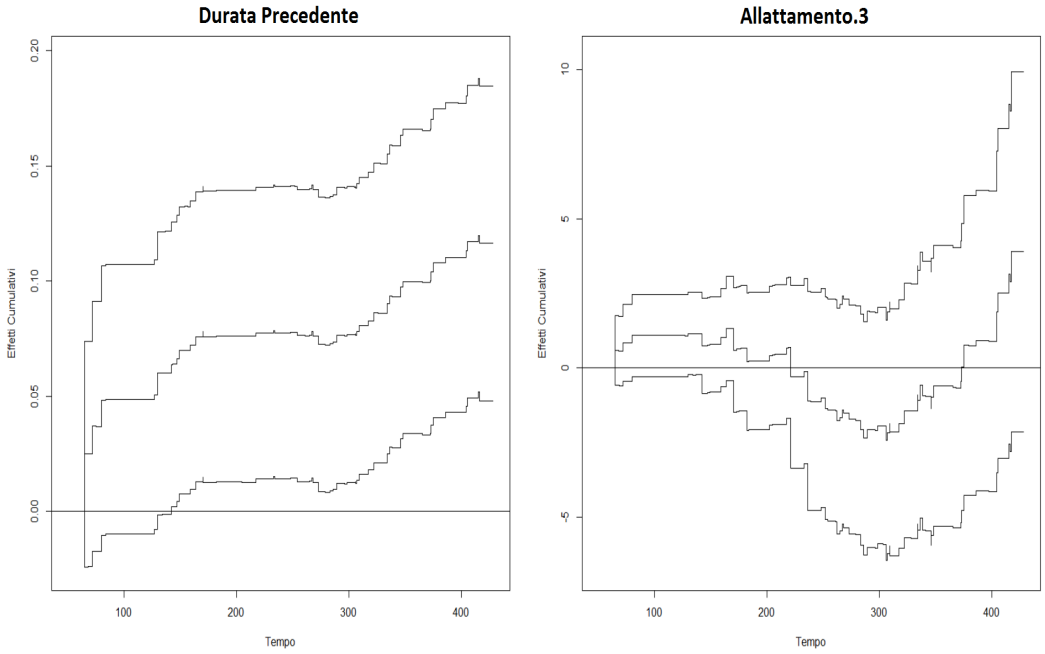


Figura 5.5: Grafico degli effetti cumulati relativi al modello semiparametrico a rischi additivi.

crescere del tempo, l'effetto dell'allattamento totale riduca il rischio di sperimentare l'evento del ritorno alla fertilità; curiosamente, verso la fine del primo anno di età del bambino, l'allattamento sembrerebbe tornare a far crescere il rischio: quest'effetto potrebbe anche essere dovuto alla scarsa numerosità campionaria di soggetti che superano il primo anno di amenorrea. Notiamo dal grafico relativo alla durata trascorsa nello stato precedente, invece, come l'effetto di questa variabile aumenti notevolmente il rischio di sperimentare il ritorno alla fertilità nei primi 150 giorni circa dopo l'ingresso nello stato relativo alla Transizione 3. L'effetto sembra completamente annullarsi in un periodo compreso tra i 150 e i 300 giorni circa, per poi ritornare ad influire positivamente sul rischio di ritorno della fertilità.

Questi sono i risultati che abbiamo ottenuto dall'analisi del secondo dei problemi di interesse. Come già ribadito in precedenza, per ottenere risultati più affidabili in studi successivi è raccomandabile aumentare la numerosità campionaria e cercare di codificare certe variabili in maniera più informativa.

Capitolo 6

Conclusioni

In questa tesi di laurea abbiamo effettuato un'analisi statistica riguardante il problema del ritorno alla fertilità nella donna dopo il parto. L'interesse dei ricercatori era volto all'individuazione dei fattori che influenzano il ritorno alla fertilità, in particolar modo è stato di interesse comprendere quale effetto abbia l'allattamento esclusivo al seno al netto dei fattori fisiologici della donna, inclusi i fattori ormonali. Nessuno studio precedente, infatti, aveva valutato l'effetto della variabile relativa all'allattamento esclusivo sul ritorno della fertilità al netto delle variabili fisiologiche e dell'attività degli ormoni $E3G$ ed LH. Abbiamo confermato il fatto che l'allattamento esclusivo al seno allunghi la durata della fase di amenorrea post-parto, anche al netto dei fattori ormonali. Abbiamo osservato come l'assenza di rilevazione di valori elevati dei due ormoni sopracitati causi un allungamento dei tempi di ritorno alla fertilità, mentre l'osservazione di questi due fattori aumenti notevolmente il rischio di sperimentare in breve tempo una ovulazione fertile con relativa mestruazione. I fattori fisiologici relativi ai soggetti sono stati raccolti al momento dell'ingresso nello studio, solamente il peso della donna si è rilevato significativo: il suo effetto aumenta il rischio di ritorno alla fertilità. In futuro, sarebbe di interesse effettuare un nuovo studio per confermare i risultati da noi ottenuti. Questa volta, sarebbe ottimale raccogliere le informazioni relative all'interruzione dell'allattamento sotto forma di data, sarebbe inoltre auspicabile raccogliere la variabile relative all'indice di massa corporea sotto codifica di variazione mensile. La numerosità campionaria rilevata nel nostro studio è risultata piuttosto esigua: uno studio futuro dovrebbe avvalersi, se possibile, di una numerosità campionaria ben più elevata al fine di ottenere una maggiore solidità dei risultati.

Abbiamo anche investigato il ruolo che le variabili esplicative gio-

cano sul fatto che si verifichi una produzione di ormone luteinizzante sufficientemente alta da essere rilevata dal dispositivo di monitoraggio elettronico. Il fatto di non rilevare alcun picco di ormone LH durante i protocolli di pianificazione familiare può creare false sicurezze nelle donne intenzionate a non sperimentare nuove gravidanze a breve termine. Sembrerebbe che al calare dell'indice di massa corporea questa osservazione risulti più difficile. Inoltre, una volta osservata una elevata quantità di ormone E3G, sembra molto più plausibile l'eventualità di osservare un picco di ormone luteinizzante.

Nel dataset ci sono state fornite anche le durate dei cicli mestruali immediatamente successivi al ritorno della fertilità nelle donne facenti parte dello studio. Potrebbe essere interessante vedere se ci siano variabili che influiscono su queste durate, ed in caso affermativo, individuare quali ed in quale misura. Dei semplici modelli a rischi proporzionali e multi-stato sono stati adattati a questo ulteriore problema, ma non sono stati in grado di rilevare alcuna variabile significativa.

Bibliografia

Aalen O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pp. 1–25. Springer.

Akaike H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Andersen P. K.; Keiding N. (2002). Multi-state models for event history analysis. *Statistical methods in medical research*, **11**(2), 91–115.

Andersen P. K.; Hansen L. S.; Keiding N. (1991). Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics*, pp. 153–167.

Andersen P. K.; Abildstrom S. Z.; Rosthøj S. (2002). Competing risks as a multi-state model. *Statistical methods in medical research*, **11**(2), 203–215.

Andridge R. R.; Little R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, **78**(1), 40–64.

Behre H.; Kuhlage J.; Gaβner C.; Sonntag B.; Schem C.; Schneider H.; Nieschlag E. (2000). Prediction of ovulation by urinary hormone measurements with the home use clearplan® fertility monitor: comparison with transvaginal ultrasound scans and serum hormone measurements. *Human Reproduction*, **15**(12), 2478–2482.

Bouchard T.; Fehring R. J.; Schneider M. (2013). Efficacy of a new postpartum transition protocol for avoiding pregnancy. *The Journal of the American Board of Family Medicine*, **26**(1), 35–44.

Breiman L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics*, pp. 89–99.

Cox D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, **34**, 187–220.

Cox D. R. (1975). Partial likelihood. *Biometrika*, pp. 269–276.

de Wreede L. C.; Fiocco M.; Putter H. *et al.* (2011). mstate: an r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, **38**(7), 1–30.

Fehring R. J.; Schneider M.; Raviele K.; Barron M. L. (2007). Efficacy of cervical mucus observations plus electronic hormonal fertility monitoring as a method of natural family planning. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, **36**(2), 152–160.

Fehring R. J.; Schneider M.; Barron M. L. (2008). Efficacy of the marquette method of natural family planning. *MCN: The American Journal of Maternal/Child Nursing*, **33**(6), 348–354.

Fine J. P.; Gray R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, **94**(446), 496–509.

Fix E.; Neyman J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, **23**(3), 205–241.

Gail M. (1975). A review and critique of some models used in competing risk analysis. *Biometrics*, pp. 209–222.

Garson G. (2015). Missing values analysis and data imputation. *North Carolina State University. Asheboro, USA: Statistical Associates Publishers*.

Gartner L. M.; Morton J.; Lawrence R. A.; Naylor A. J.; O'Hare D.; Schanler R. J.; Eidelman A. I. (2005). Breastfeeding and the use of human milk. *Pediatrics*, **115**(2), 496–506.

Grambsch P. M.; Therneau T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, pp. 515–526.

Gray R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pp. 1141–1154.

Hanley J. A. (2008). The breslow estimator of the nonparametric baseline survivor function in cox's regression model: some heuristics. *Epidemiology*, **19**(1), 101–102.

Hougaard P. (1999). Multi-state models: a review. *Lifetime data analysis*, **5**(3), 239–264.

Hougaard P. (2012). *Analysis of multivariate survival data.* Springer Science & Business Media.

Huberman M.; Langholz B. (1999). Application of the missing-indicator method in matched case-control studies with incomplete data. *American journal of epidemiology*, **150**(12), 1340–1345.

İnce N.; Özyildirim B.; Işik E.; Bozcali F. (2007). The effect of breast-feeding in contraception which is a method of natural family planning. *Iranian Journal of Public Health*, **36**(4), 12–19.

Ishwaran H.; Kogalur U. B.; Blackstone E. H.; Lauer M. S. (2008). Random survival forests. *The annals of applied statistics*, pp. 841–860.

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

Kay R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, pp. 227–237.

Kennedy K. I.; Rivera R.; McNeilly A. S. (1989). Consensus statement on the use of breastfeeding as a family planning method. *Contraception*, **39**(5), 477–496.

Klein J. P.; Moeschberger M. L. (2005). *Survival analysis: techniques for censored and truncated data.* Springer Science & Business Media.

Lando D.; Skødeberg T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *journal of banking & finance*, **26**(2), 423–444.

Martinussen T.; Scheike T. H. (2007). *Dynamic regression models for survival data.* Springer Science & Business Media.

McCann M. F.; Liskin L.; Piotrow P. T.; Rinehart W.; Fox G. (1981). Breast-feeding fertility and family planning. *Population Reports. Series J: Family Planning Programs*, pp. 1–51.

McKeague I. W.; Sasieni P. D. (1994). A partly parametric additive risk model. *Biometrika*, **81**(3), 501–514.

McNeilly A. S. (2002). Lactational control of reproduction. *Reproduction, Fertility and Development*, **13**(8), 583–590.

Meira-Machado L. F.; de Uña-Álvarez J.; Cadarso-Suárez C.; Andersen P. (2008). Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research.*

Pepe M. S.; Mori M. (1993). Kaplan meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in medicine*, **12**(8), 737–751.

Perez A.; Labbok M. H.; Queenan J. T. (1992). Clinical study of the lactational amenorrhoea method for family planning. *The Lancet*, **339**(8799), 968–970.

Prentice R. L.; Kalbfleisch J. D.; Peterson Jr A. V.; Flournoy N.; Farewell V.; Breslow N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, pp. 541–554.

Putter H. (2016). Tutorial in biostatistics: Competing risks and multi-state models analyses using the mstate package. *Statistics in medicine*.

Putter H.; van der Hage J.; de Bock G. H.; Elgalta R.; van de Velde C. J. (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical journal*, **48**(3), 366–380.

Putter H.; Fiocco M.; Geskus R. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, **26**(11), 2389–2430.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins J. M.; Rotnitzky A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *Aids Epidemiology*, pp. 297–331. Springer.

Robinson J. E.; Wakelin M.; Ellis J. E. (2007). Increased pregnancy rate with use of the clearblue easy fertility monitor. *Fertility and sterility*, **87**(2), 329–334.

Royston P. (1991). Identifying the fertile phase of the human menstrual cycle. *Statistics in Medicine*, **10**(2), 221–240.

Royston P. *et al.* (2004). Multiple imputation of missing values. *Stata journal*, **4**(3), 227–41.

Rubin D. B.; Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**(394), 366–374.

Schafer J. L.; Graham J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, **7**(2), 147.

Shah A. D.; Bartlett J. W.; Carpenter J.; Nicholas O.; Hemingway H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, **179**(6), 764–774.

Städler N.; Bühlmann P. (2010). Pattern alternating maximization algorithm for high-dimensional missing data. *Arxiv preprint arXiv*, **1005**.

Stekhoven D. J. (2015). Missforest: nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*, **1**, 05011.

Stekhoven D. J.; Bühlmann P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**(1), 112–118.

Sverdrup E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Scandinavian Actuarial Journal*, **1965**(3-4), 184–211.

Therneau T. M.; Grambsch P. M. (2013). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Therneau T. M.; Grambsch P. M.; Fleming T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, pp. 147–160.

Triantaphyllou E. (2000). Multi-criteria decision making methods. In *Multi-criteria Decision Making Methods: A Comparative Study*, pp. 5–21. Springer.

Troyanskaya O.; Cantor M.; Sherlock G.; Brown P.; Hastie T.; Tibshirani R.; Botstein D.; Altman R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**(6), 520–525.

Valeggia C.; Ellison P. T. (2009). Interactions between metabolic and reproductive functions in the resumption of postpartum fecundity. *American Journal of Human Biology*, **21**(4), 559–566.

Van Buuren S.; Oudshoorn K. (1999). *Flexible mutlivariate imputation by MICE*. Leiden: TNO.

van der Heijden G. J.; Donders A. R. T.; Stijnen T.; Moons K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*, **59**(10), 1102–1109.

Victora C. G.; Bahl R.; Barros A. J.; França G. V.; Horton S.; Krasevec J.; Murch S.; Sankar M. J.; Walker N.; Rollins N. C. *et al.* (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *The Lancet*, **387**(10017), 475–490.

Waljee A. K.; Mukherjee A.; Singal A. G.; Zhang Y.; Warren J.; Balis U.; Marrero J.; Zhu J.; Higgins P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, **3**(8), e002847.

White I. R.; Royston P.; Wood A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, **30**(4), 377–399.

WorldHealthOrganization (1981). Temporal relationships between ovulation and defined changes in the concentration of plasma estradiol-17β, luteinizing hormone, follicle-stimulating hormone, and progesterone: Ii. histologic dating. *American Journal of Obstetrics and Gynecology*, **139**(8), 886–895.

Xu J. (1999). Survival analysis: Techniques for censored and truncated data (statistics for biology and health). *American Journal of Epidemiology*, **149**(2), 200.