Modellazione statistica delle donazioni di sangue nella provincia di Trieste

Invalid Date

La ricerca studia i registri dei donatori di sangue di Trieste (2009-2023) per stimare le caratteristiche dei donatori e predire il loro comportamento futuro.

Su questo pannello longitudinale vengono applicati:

(i) modelli di regressione quasi-Poisson, tweedie e Gamma per spiegare il numero totale di donazioni effettuate da ciascun individuo;

(ii) Hidden Markov Models con covariate trattate in modo Bayesiano per descrivere traiettorie latenti di comportamento donativo, e studiare come le covariate influenzino tali stati latenti.

UNIVERSITÀ DEGLI STUDI DI TRIESTE

DIPARTIMENTO DI SCIENZE ECONOMICHE, AZIENDALI, MATEMATICHE E STATISTICHE "BRUNO DE FINETTI"

Laurea Magistrale in Scienze Statistiche Attuariali

Hidden Markov Models e Generalized Lienar Models Applicati alle Donazioni di Sangue

Relatore:

Prof: Leonardo Egidi

Candidato:

Erik De Luca

ANNO ACCADEMICO 2024/2025

Indice

1 Blood Donation Predictor **11**

Prefazione **13**

2 Introduzione **15**
2.1 I Dati . . . 16
2.1.1 Fonte ed Elaborazione . . . 18
2.1.2 Analisi Preliminare . . . 20
2.2 Integrazione dei Dati . . . 20
2.2.1 Stima dei Residenti Passati . . . 21
2.2.2 Unione dei Dati . . . 22

3 Modelli Lineari Generalizzati **25**
3.1 Teoria dei GLM . . . 25
3.1.1 Regressione Lineare . . . 25
3.1.2 Metodi di stima . . . 26
3.1.3 Estensione della Regressione Lineare . . . 26
3.1.4 Famiglia Esponenziale . . . 27
3.2 Modello Esplicativo . . . 28
3.2.1 Quasi-Poisson . . . 28
3.2.2 Tweedie *(power $\sim$ 1.19)* . . . 30
3.2.3 Gamma . . . 31
3.3 Modello Predittivo . . . 33
3.3.1 Effetto Covid . . . 33
3.3.2 Modello finale . . . 33

4 Bayesian Hidden Markov Models **35**
4.1 Accenni di Teoria . . . 35
4.1.1 Processo Markoviano . . . 35
4.1.2 Hidden Markov Models (HMM) . . . 37
4.1.3 Bayesiana . . . 38
4.2 Il modello applicato alle donazioni . . . 38
4.2.1 Pyro . . . 38
4.2.2 Dati e covariate . . . 39
4.2.3 Componenti del Modello . . . 42
4.3 Risultati . . . 48
4.4 Algoritmi e diagnostiche . . . 48
4.4.1 Viterbi con covariate . . . 48
4.4.2 Forward per log-likelihood e previsione . . . 48

4.4.3 Occupancy e switch-rate 48
4.4.4 Calibrazione e confronto con GLM 49
4.5 Scelta del numero di stati 49
4.6 Riproducibilità e note operative 49

5 Dashboard e Sito Web **51**
5.1 Motivazione 51
5.2 Sviluppo 51

6 Conclusioni **53**
6.1 Risultati 53
6.2 Idee per il Futuro 53

Riferimenti **55**

Elenco delle Figure

2.1 Distribuzioni dei donatori per genere eđetà nel 2009 e nel 2023 20
2.2 Distribuzione del tasso di donatori tra i residenti per anno, età e genere 23

4.1 Risultati principali del modello 47

1 Blood Donation Predictor

La ricerca studia i registri dei donatori di sangue di Trieste (2009-2023) per stimare le caratteristiche dei donatori e predire il loro comportamento futuro.

Su questo pannello longitudinale vengono applicati:

1) modelli di regressione quasi-Poisson, tweedie e Gamma per spiegare il numero totale di donazioni effettuate da ciascun individuo;
2) Hidden Markov Models con covariate trattate in modo Bayesiano per descrivere traiettorie latenti di comportamento donativo, e studiare come le covariate influenzino tali stati latenti.

Prefazione

Avviso

Not ready yet. Come back at the end of september

2 Introduzione

L'obiettivo di questa tesi è proporre — e validare empiricamente — un framework statistico per la previsione delle donazioni di sangue nel territorio Giuliano-Isontino, con orizzonte annuale.

La scelta di concentrare l'analisi sulla provincia di Trieste è motivata da tre fattori:

1. **Alta densità di donatori rispetto alla popolazione residente**: il territorio triestino è storicamente virtuoso nella raccolta di sangue; ciò rende disponibili serie temporali lunghe e relativamente complete.
2. **Stabilità del bacino d'utenza**: la popolazione residente ha oscillazioni demografiche contenute, riducendo la variabilità "esterna" dovuta a migrazioni massicce. A differenza di grandi città, come Milano e Roma dove i flussi migratori sono hanno un impatto maggiore.
3. **Accesso a dati granulari**: l'ASUGI ha messo a disposizione un estratto anonimo dei registri donatori 2009-2023 che, pur privo di variabili sensibili, contiene informazioni anagrafiche e cronologia donativa sufficienti per la modellazione.

Nella pratica quotidiana i centri trasfusionali devono rispondere alla domanda clinica garantendo un buffer di scorte: sovrastimare i lotti in scadenza comporta costi di smaltimento, mentre sottostimare la raccolta può generare situazioni critiche con rinvii di interventi chirurgici programmati.

Pertanto, è cruciale disporre di **strumenti predittivi** che stimino:

- la probabilità che un donatore torni a donare l'anno successivo;
- la distribuzione del numero di donazioni attese per singolo individuo;
- i profili latenti di comportamento (frequente, saltuario, non-donatore, ecc.) e la loro evoluzione nel tempo.

A fronte di tali esigenze la tesi si articola in due nuclei metodologici:

1. **Generalized Linear Models** (GLM) per stimare il numero *cumulato* di donazioni nell'orizzonte storico, con diverse famiglie di distribuzione (quasi-Poisson, Tweedie, Gamma).

2. **Hidden Markov Models bayesiani** con covariate, in cui il numero di donazioni annue è trattato come emissione di una variabile di stato latente.

   Il fitting è condotto mediante Variational Inference (Pyro), consentendo di gestire milioni di osservazioni con tempi computazionali compatibili al problema.

A corredo dei modelli è stata sviluppata una **dashboard interattiva Quarto + Shiny** che consente al personale medico di:

- filtrare la popolazione (età, genere, anno di prima donazione, ecc.);
- visualizzare la probabilità di transizione fra stati latenti;

2.1 I Dati

In Italia possono donare sangue le persone di età compresa tra i 18 e i 65 anni, con un peso corporeo superiore ai 50 kg e in buono stato di salute. Gli uomini e le donne non in età fertile possono donare sangue intero ogni 3 mesi, mentre le donne in età fertile possono farlo 2 volte l'anno.

La donazione di sangue nel contesto italiano è il frutto di un sistema dove stakeholder del settore pubblico (regioni, centri ospedalieri), privato (associazioni non profit) e cittadini contribuiscono attivamente al buon esito di questo processo (Guglielmetti Mugion et al., 2021). La creazione del Centro Nazionale Sangue (CNS) e del Registro nazionale del sangue nel 2007 ha trasformato l'assetto organizzativo della donazione del sangue in Italia. Il CNS è stato istituito con Decreto del Ministro della Salute del 26 aprile 2007 e ha iniziato il suo mandato il 1° agosto dello stesso anno. Il CNS svolge funzioni di coordinamento e controllo tecnico-scientifico del sistema trasfusionale nazionale nelle materie disciplinate dalla legge n. 219 del 21 ottobre 2005 "Nuova disciplina delle attività trasfusionali e della produzione nazionale degli emoderivati" e dai decreti di trasposizione delle direttive europee. All'interno di tale sistema sono presenti le Strutture Regionali di Coordinamento per le attività trasfusionali (SRC). Le SRC sono strutture tecnico-organizzative delle Regioni e Province Autonome che garantiscono il supporto alla programmazione nazionale in materia di attività trasfusionali e il coordinamento e controllo tecnico-scientifico della rete trasfusionale regionale, in sinergia con il Centro Nazionale Sangue. Queste strutture regionali, anche definite Centri Regionali Sangue, detengono la responsabilità della raccolta e gestione delle donazioni di sangue a livello regionale.

Nel corso degli anni, il Ministero della Salute ha visto un significativo supporto dalle associazioni attive nel campo delle donazioni, come AVIS (Associazioni Volontari Italiani Sangue), FRATRES (Consociazione Nazionale dei Gruppi Donatori di Sangue Fratres delle Misericordie d'Italia), FIDAS (Federazione Italiana Associazioni Donatori di Sangue), Croce Rossa Italiana. Queste organizzazioni svolgono un ruolo cruciale nel promuovere attivamente la pratica della donazione del sangue: sebbene la decisione di donare sia una scelta individuale, è infatti importante sottolineare

il ruolo essenziale che esse svolgono nell'informare e nel fungere da ponte tra le istituzioni (scuole incluse) e i cittadini. La pratica del dono del sangue, peraltro, produce capitale sociale non solo per il dono di una parte di sé in quanto tale, ma anche grazie alla partecipazione sociale all'interno delle organizzazioni che la promuovono.

La letteratura riguardante il dono nel sangue nel contesto italiano si è concentrata principalmente sui donatori e sulle motivazioni al dono. Molte delle ricerche sono state svolte in collaborazione con l'AVIS, probabilmente perché la raccolta di informazioni riguardanti le unità di sangue donate si è sistematizzata a livello nazionale solo dal 2007 con l'istituzione del CNS e la creazione del registro nazionale sangue, come descritto poc'anzi. Le ricerche hanno sovente utilizzato dati raccolti intervistando i donatori (sia attraverso strumenti standardizzati, sia con approcci di tipo qualitativo). Nei prossimi paragrafi utilizzeremo i dati sulle donazioni di sangue a livello territoriale, ma prima di entrare nel dettaglio sulle differenze geografiche delle donazioni di sangue pare opportuna una sintetica ricognizione sui principali aspetti emergenti da tali ricerche condotte nel contesto italiano.

Un'analisi approfondita condotta da Lacetera e Macis (2013) sui dati AVIS relativi al periodo 1983-2006, focalizzata su una città del centro-nord Italia, ha quantificato l'impatto della Legge 584 del 1967, che ha stabilito il riconoscimento del diritto a una giornata di riposo dal lavoro e alla piena retribuzione al donatore di sangue, sulle pratiche di donazione. Lo studio ha rilevato che l'applicazione di tale normativa ha indotto i donatori a effettuare, in media, una donazione aggiuntiva all'anno. Attraverso un'analisi comparativa delle frequenze di donazione associate ai diversi stati occupazionali assunti dal medesimo individuo, i ricercatori hanno evidenziato una correlazione significativa tra l'occupazione e la propensione alla donazione.

I risultati hanno dimostrato che, in media, quando un individuo è occupato e quindi idoneo a beneficiare dell'incentivo del giorno di riposo retribuito, la frequenza annuale delle donazioni aumenta di circa un'unità rispetto ai periodi di non occupazione.

La decisione di donare sangue è fortemente influenzata da fattori personali e sociali. Una ricerca condotta a Bergamo nel 2006 (Bani, Strepparava, 2011) ha mostrato che il 50% dei donatori è stato motivato dal confronto con amici e familiari, ma anche l'aver ricevuto trasfusioni o conoscere qualcuno che ha beneficiato di una trasfusione hanno avuto un impatto significativo, aumentando la frequenza delle donazioni e la propensione a persuadere altri a donare. Questi risultati evidenziano l'importanza delle relazioni personali e delle esperienze dirette nella promozione della donazione di sangue, confermando il forte legame emotivo e sociale che spinge le persone a donare.

Anche il lavoro delle associazioni, come ricordato, è fondamentale. Esse contribuiscono all'incremento di capitale sociale sia costruendo relazioni con le istituzioni locali, sia creando partecipazione e attivazione attraverso diverse iniziative che mirano allo sviluppo di senso di comunità e appartenenza (Saturni, 2013).

I giovani che partecipano all'AVIS (Bassi et al., 2024), percepiscono l'associazione non solo come un'opportunità per donare il sangue, ma anche come un punto di

riferimento fondamentale per la comunità. All'interno della ricerca di Bassi e colleghi gli intervistati sottolineano il ruolo dell'AVIS come infrastruttura sociale capace di promuovere la coesione e l'inclusione. Essi vedono nel volontariato un'occasione per tessere relazioni, costruire reti e contribuire attivamente alla vita della comunità, confermando così l'importanza delle associazioni come presidi territoriali e promotori del capitale sociale.(Bordandini 2025)

2.1.1 Fonte ed Elaborazione

Tabella 2.1: Variabili presenti nel dataset fornito dal centro trasfusionale

campo	descrizione
`donor_class`	classificazione del donatore
`donation_type`	categoria (SANGUE, PLASMA, PIASTRINE, ...)
`birth_year`	anno di nascita
`birth_cohort`	coorte di nascita, generazione (1970, 1975, 1980, ...)
`first_donation_year`	anno della prima donazione registrata
`first_donation_cohort`	coorte della prima donazione registrata
`number of donations`	numero di donazioni effettuate nel determinato anno
`gender`	M/F
`year`	anno di riferimento delle donazioni
`age`	età del donatore
`unique_number`	identificativo anonimo del donatore

I dati provengono dall'estrazione e anonimizzazione di dati provenienti dal data warehouse dell'Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI). I dati vengono forniti in formato Il dataset primario proviene dal sistema informativo dei centri trasfusionali dell'ASUGI e contiene le donazioni fatte da un individuo in un determinato anno, corredate di ulteriori informaizoni, riportate nella Tabella 2.1.

donor class	dona...	type	birth year	birt...	hon	firs...	year	firs...	hon	numb...	gen	gender	year	age	uniq...	mber
P	SANGUE		1984	1980		2002		2000		1	F		2023	39	26771385	
P	PLASMAF		1976	1975		2003		2000		6	M		2018	42	26790414	
P	SANGUE		1966	1965		2009		2005		2	M		2020	54	26940369	
P	AFPLTPLA		1996	1995		2014		2010		1	F		2019	23	27058047	
P	PLASMAF		1989	1985		2009		2005		2	F		2023	34	26926947	
P	AFPLTPLA		1961	1960		1985		1985		1	M		2011	51	26514894	
P	PLTAFE		1962	1960		1993		1990		2	M		2014	52	26490138	
P	PLTAFE		1970	1970		2001		2000		3	F		2016	46	26681457	

Tabella 2.2: Campione di dati grezzi forniti dal centro trasfusionale

Anche se i dati provengono da un datawarehouse che rispetta protocolli nella gestione del dato, è comunque necessaria una fase di *ETL* (Extract, Transform, Load). Ovvero l'estrazione e la trasformazione del dato in modo da adattarlo e arrichirlo ai nostri scopi specifici. Il lavoro viene svolto mediante il linguaggio di programmazione statistica R, e la libreria `tidyverse`, una libreria contente funzioni per la manipolazione dei dati garantendo leggibilità del codice e velocità d'esecuizione.

La tabella originale è composta da 268.530 righe, una per ogni donatore e per ogni anno nella quale abbia donato. Tuttavia sono presenti record duplicati (171.378).

I passi principali compiuti sono i seguenti:

- rimozione record duplicati (`janitor::get_dupes`);
- sostituzione di `NA` con 0 nei conteggi annuali;
- derivazione di età = `year - birth_year` e classi d'età quinquennali;
- standardizzazione (`z`-score) di `birth_year` e `age` per facilitare la convergenza dell'ottimizzatore nei modelli bayesiani;
- aggiunta della variabile dummy Covid;
- creazione di **tre matrici di covariate**:
  - x^{π} (fisse per donatore): anno di nascita e genere;
  - x_t^A (tempo-varianti): età categorica e dummy Covid;
  - x_t^{em} (tempo-varianti): età categorica, dummy Covid e genere.

La derivazione dell'età sarà utile per avere una variabile dinamica, che varia nel tempo. Infatti si presuppone che la propensione al donare vari con l'età del donatore, ed avendo osservazioni pluriennali, si ritiene opportuno tenere in considerazione ciò. Però, anche il fattore generazionale può influire, ovvero una persona di 50 anni del 1970 può avere una propensione nel donare diversa di un cinquantenne nato 10 anni prima. Questo potrebbe essere dovuto da fattori generazionali e anch'esso andrà incluso nelle analisi. Infine, in molte analisi condotte, è stata utilizzata la variabile età come categoriale, anziché numerica.

Nei primi 10 giorni di marzo 2020, all'inzio della pandemia SARS-CoV-2, le donazioni di sangue in Italia sono state quasi nulle, per poi passare ad un forte aumento. Il Centro Nazionale Sangue (CNS), l'autorità nazionale competente, pubblicò linee guida chiare per permettere la continuazione dei prelievi di sangue ed evitare un'interruizione della catena di approvigionamento della raccolta di materiale trasfusionale. (Pati et al. 2021)

Infine, le osservazioni vengono aumentate, ossia vengono aggiunte le donazioni pari a 0 negli anni in cui non abbiamo il dato di uno specifico donatore. Si ipotizza che quando il dato non venga raccolto non ci siano donazioni da parte dell'individuo. Questa è un'ipotesi forte, infatti ci potrebbero essere diverse ragioni per la mancanza del dato e non solo la mancata donazione. Ad esempio, l'individuo si sarebbe potuto

trasferire, o avrebbe potuto decidere di donare in un centro trasfusionale differente da quello da noi analizzato, ossia il centro trasfusionale dell'ASUGI.

2.1.2 Analisi Preliminare

L'Italia, come molti altri paesi del mondo, è affetta dal fenomeno generazionale del *baby boom*, ovvero da un notevole incremento delle nascite negli anni 50-60, dovuto a vari fattori, tra cui la forte crescita economica. Questo fenomeno è oggetto di studio in diversi ambiti, uno tra i quali è l'ambito assicurativo, dove ci si preoccupa se le generazioni più giovani saranno in grado di sopportare il sistema pensionistico quando i *baby boomers* andranno in pensione. Lo stesso discorso vale per le donazioni di sangue, in quanto i donatori saranno meno e coloro che necesiteranno di sacche di sangue sarà sempre maggiore. Questo fenomeno generazionale si può osservare dalla Figura 2.1 dove si osserva unon spostamento nella "gobba" verso l'alto dal 2009 al 2023. Le donazioni provengo maggiormente dagli uomini, tuttavia si osserva (vedi Tabella 2.3) come questo gap si sti riducendo nelle nuove generazioni, grazie anche alla efficace comunicazione dei centri di trasfusione.

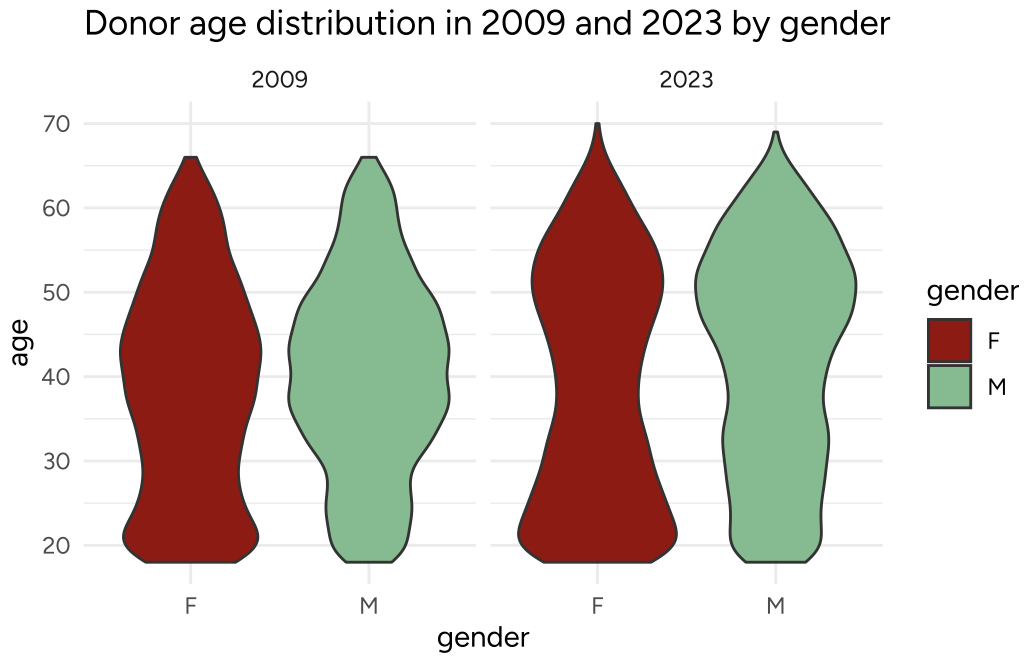


Figura 2.1: Distribuzioni dei donatori per genere eđetà nel 2009 e nel 2023

2.2 Integrazione dei Dati

Le analisi eseguite finora sono state condotte sui donatori e le loro donazioni, mostrandoci caratteristiche e pattern fondamentali sulle donazioni. Tuttavia, per un'analisi più approfondita è necessario tenere in considerazione anche quella parte

		2009	2023
F	(20,30]	802	866
	(30,40]	872	590
	(40,50]	962	713
	(10,20]	341	398
	(50,60]	574	729
	(60,70]	136	175
M	(20,30]	1286	1142
	(30,40]	2275	1249
	(50,60]	1164	1685
	(40,50]	2362	1649
	(60,70]	350	379
	(10,20]	358	382

Tabella 2.3: Donazioni nel 2009 e 2023 per fasce d'età e per genere

di popolazione che non dona, e che, di conseguenza, non risulta essere presente nei dati a noi disponibili.

L'obiettivo è di integrare il database che possediamo con ulteriori informazioni che potrebbero arrichire le analisi e aggiungere informazioni ai modelli.

I dati in nostro possesso sono la raccolta delle donazioni presso le strutture sanitarie dell'ASUGI, ovvero dell'Azienda Sanitaria Universitaria Giuliano Isontina, ovvero che i dati in nostro possesso provengono dai centri trasfusionali del territorio di Trieste e Gorizia. Tuttavia, ciò non indica che le donazioni provengano da cittadini residenti nel territorio Giuliano-Isontino. Infatti, le donazioni sono aperte a tutti, anche a cittadini stranieri, come potrebbe essere uno studente durante il suo progetto Erasmus a Trieste. Andranno fatto, quindi, delle assunzioni per semplificare la realtà. Si ipotizza che le donazioni provengono solo da residenti del territorio. Possiamo allora integrare i dati con le informazioni sui residenti.

Le informazioni provengono dal database pubblico dell'Istituto Nazionale di Statistica, ISTAT. I dati in formato tabulare contengono informazioni di vario genere, tra cui il genere, l'anno, la popolazione e lo stato civile. I dati vengono quindi processati e adattati ai dati sulle donazioni di sangue.

2.2.1 Stima dei Residenti Passati

L'ISTAT diffonde serie complete per il 2019–2023, mentre il nostro dataset copre il periodo a partire dal 2009. Per ricostruire a ritroso la popolazione residente nel capoluogo giuliano adottiamo un approccio in due passi. Primo, richiamamo l'identità di bilancio demografico, che regola l'evoluzione della popolazione residente (cfr. {ISTAT (2023a)}):

$$P_t = P_{t-1} + N_t - M_t + I_t - E_t$$

dove P_t è la popolazione a fine anno t, N_t i nati vivi, M_t i decessi, I_t gli iscritti per migrazione ed E_t i cancellati per migrazione.

Per ricostruire a ritroso la popolazione residente nel capoluogo giuliano (2009–2018) adottiamo il metodo di retroproiezione per coorti ("reverse life-table"), che utilizza i sopravviventi l_x dalle tavole di mortalità per risalire alla consistenza delle coorti alle età precedenti (cfr. {Caselli, Vallin, e Wunsch (2006)}; per definizioni e stima di l_x nelle tavole ISTAT si veda {ISTAT (2023b)}). La formulazione operativa impiegata è:

$$n_x^{y_i} = n_{x-(y_i-y_j)}^{y_j} \frac{l_x}{l_{x-(y_i-y_j)}},$$

dove $n_x^{y_i}$ è l'effettivo alla età x al tempo y_i, $n_{x-(y_i-y_j)}^{y_j}$ è l'effettivo della medesima coorte alla età $x-(y_i-y_j)$ osservato (o stimato) al tempo y_j, e il rapporto $\frac{l_x}{l_{x-(y_i-y_j)}}$ rappresenta la probabilità di sopravvivenza tra le due età secondo la tavola di mortalità di riferimento. In mancanza di flussi migratori comunali completi, si assume, in prima approssimazione, migrazione netta nulla o costante e si verifica la robustezza dei risultati tramite analisi di sensitività.

> **i** Nota sulle tavole SIM/SIF 02
>
> Con "SIM/SIF 02" si fa riferimento a una delle serie ufficiali di tavole di mortalità pubblicate da ISTAT, per età singola e per sesso, che riportano le principali funzioni di tavola (in particolare i sopravviventi l_x, le probabilità q_x, i decessi d_x e gli esposti L_x) e costituiscono la base standard per calcoli di sopravvivenza e retroproiezione a livello nazionale.

2.2.2 Unione dei Dati

La join tra il registro dei donatori e il dataframe con i residenti consente di costruire un indicatore di "penetrazione" (quota di donatori sulla popolazione residente), disaggregato per anno, classe d'età e genere. Indichiamo con $\#\text{donatori}(a, y, g)$ il numero di individui presenti nel dataset in quella cella (a, y, g) e con $\text{residenti}(a, y, g)$ il denominatore demografico coerente (stessa cella di età, anno e genere). Definiamo quindi:

$$\text{penetration}_{a,y,g} = \frac{\#\text{donatori}(a, y, g)}{\text{residenti}(a, y, g)}, \qquad g \in \{F, M\},\ a = \text{classe d'età},\ y.$$

class_age	2011	2014	2017	2020	2023
(10,20]	7.48%	7.63%	7.76%	6.28%	7.81%
(20,30]	5.06%	4.92%	5.04%	5.25%	5.08%
(30,40]	4.45%	4.01%	4.06%	3.81%	4.01%
(40,50]	4.16%	3.66%	3.87%	3.69%	3.69%
(50,60]	2.94%	2.69%	2.93%	3.03%	3.03%
(60,70]	1.06%	0.83%	0.80%	0.87%	1.01%

Tabella 2.4: Percentuale di donatori tra i residenti per età e anno

Nel seguito, per il grafico (Figura 2.2) lavoriamo su età singola (poi mostrata in classi nei pannelli), mentre per la tabella (Tabella 2.4) aggreghiamo a classi decennali. Si noti che la coerenza del denominatore è garantita dall'integrazione con i residenti ISTAT per la medesima triade (g, y, a).

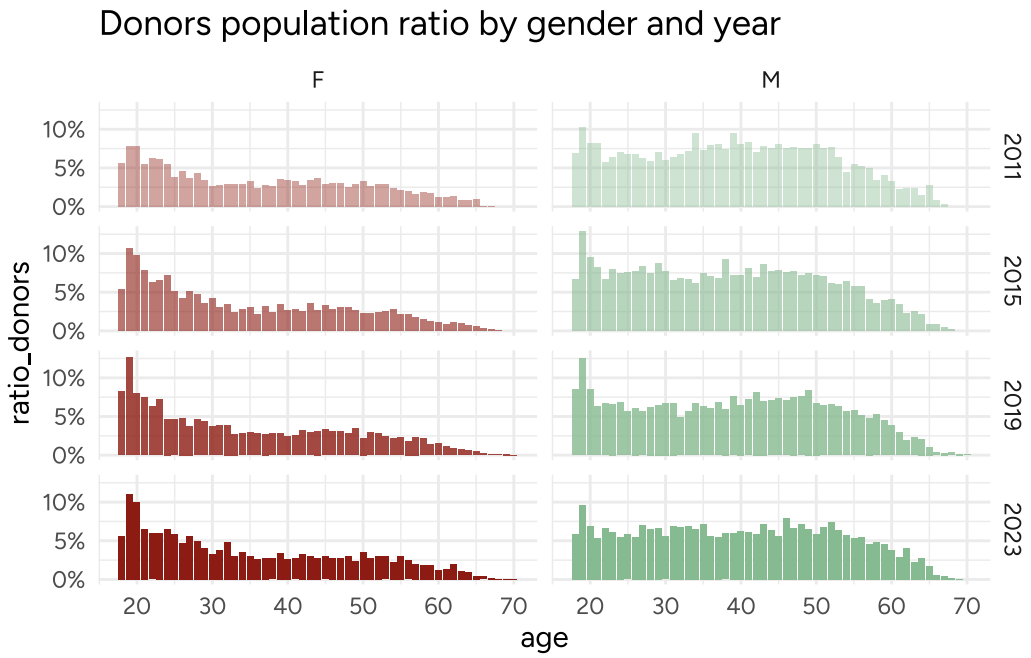


Figura 2.2: Distribuzione del tasso di donatori tra i residenti per anno, età e genere

Nel 2002, Cartocci riportava un tasso di donatori pari a 384 per 10.000 residenti; successivamente, i dati sintetizzati da Paola Bordandini (vedi Bordandini (2025)) indicano un incremento a 438 nel 2009 e a 454 nel 2022. Sulle nostre elaborazioni per Trieste — calcolate come rapporto tra donatori unici annui e residenti ISTAT nello stesso anno (per età e genere, poi aggregati) — il tasso complessivo risulta pari a 230 per 10.000 nel 2009, raggiunge un minimo di 216 nel 2021 e risale a 223 nel 2023, segnalando un lieve impatto congiunturale della pandemia e successivamente recuperato.

3 Modelli Lineari Generalizzati

3.1 Teoria dei GLM

I modelli lineari generalizzati (GLM) sono un'estensione dei modelli di regressione lineare. Per comprendere appieno il significato di Generalized Linear Model si introducono prima i modelli di regressione lineare, i metodi di stima dei coefficienti e gli elementi che li caratterizzano. Successivamente si passa alla loro estensione e all'utilizzo della famiglia esponenziale, citando alcune distribuzioni che saranno poi impiegate nei modelli.

3.1.1 Regressione Lineare

Un modello è una rappresentazione matematica di un processo che genera dati, ovvero una semplificazione della realtà basata su assunzioni probabilistiche. La regressione lineare studia la relazione tra una variabile dipendente Y e un insieme (eventualmente vuoto) di covariate X.

> **Nota**
>
> A causa della molteplicità di applicazioni in ambiti diversi, non esiste una terminologia univoca. Nel testo Y sarà anche chiamata variabile dipendente, outcome, output o target; X potrà essere detta variabile indipendente, covariata, feature, predittore o input.

La correlazione tra X e Y è misurata dalla covarianza,

$$\mathrm{Cov}(X,Y) \;=\; \mathbb{E}\Big[(X-\mathbb{E}X)(Y-\mathbb{E}Y)\Big] \;=\; \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y],$$

e dal coefficiente di correlazione

$$\rho_{XY} \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y},$$

dove σ_X e σ_Y sono le deviazioni standard. La correlazione descrive associazioni lineari, non causalità.

Quando si vuole modellare la dipendenza media di Y da X, si introduce la funzione di regressione $m(x)=\mathbb{E}[Y \mid X=x]$. Nella regressione lineare semplice,

$$\mathbb{E}[Y \mid X] \;=\; \alpha+\beta X,$$

dove α è l'intercetta e β il coefficiente di regressione. Nel caso con più covariate si impiega il predittore lineare $\eta = \alpha + X\beta$.

> **Nota**
>
> È comodo inglobare l'intercetta nella matrice delle covariate aggiungendo una colonna di 1, e scrivere il modello come
>
> $$Y \;=\; X\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2 I.$$
>
> I dettagli di stima sono rimandati alla sezione "Metodi di stima".

3.1.2 Metodi di stima

Nel modello lineare classico $Y = X\beta + \varepsilon$ con $\mathbb{E}[\varepsilon] = 0$ e $\mathrm{Var}(\varepsilon) = \sigma^2 I$, lo stimatore ai minimi quadrati ordinari (OLS) minimizza la somma dei quadrati dei residui:

$$\hat{\beta}_{\mathrm{OLS}} \;=\; \arg\min_{\beta} \, (Y - X\beta)^{\top}(Y - X\beta) \;=\; (X^{\top}X)^{-1}X^{\top}Y.$$

Se si assume in più $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, allora $\hat{\beta}_{\mathrm{OLS}}$ coincide con lo stimatore di massima verosimiglianza (MLE). La funzione di verosimiglianza è

$$L(\beta, \sigma^2; Y) \;\propto\; (\sigma^2)^{-n/2} \exp\!\left(-\frac{1}{2\sigma^2}\,(Y - X\beta)^{\top}(Y - X\beta)\right),$$

e massimizzarla rispetto a β equivale a minimizzare la somma dei quadrati.

> **Curiosità**
>
> Nei GLM la stima avviene per massima verosimiglianza nella famiglia esponenziale. Indicando con $\mu = \mathbb{E}[Y \mid X]$ e con $g(\mu) = \eta = X\beta$ la funzione di collegamento, le equazioni di score portano all'algoritmo IRLS (Iteratively Reweighted Least Squares): a ogni iterazione si risolve un problema di minimi quadrati pesati
>
> $$\beta^{(t+1)} \;=\; \arg\min_{\beta} \; \left\| W^{1/2}\,(z - X\beta) \right\|^2,$$
>
> dove W è una matrice di pesi dipendente da $\mu^{(t)}$ e z è la "variabile dipendente lavorata" (working response). La convergenza fornisce $\hat{\beta}_{\mathrm{MLE}}$.

3.1.3 Estensione della Regressione Lineare

Il GLM si compone di tre elementi:

- componente aleatoria: $Y_i \mid X_i \sim$ famiglia esponenziale(μ_i, ϕ);

- componente sistematica: $\eta_i = x_i^\top \beta$;
- funzione di collegamento: $g(\mu_i) = \eta_i$.

Il collegamento canonico rende lineare il parametro naturale (es. logit per Binomiale, log per Poisson, inversa per Gamma). Ogni famiglia è caratterizzata da una funzione di varianza $V(\mu)$ che determina la struttura di dispersione: $\mathrm{Var}(Y_i \mid X_i) = \phi\, V(\mu_i)$.

La qualità di adattamento si valuta tramite devianza, AIC/BIC e diagnostiche dei residui. La scelta di famiglia e link è guidata dalla natura della risposta (discreta/continua, supporto, presenza di zeri) e dall'interpretabilità dei coefficienti.

3.1.4 Famiglia Esponenziale

Linear indica la linearità nei parametri. Infatti, la funzione di regressione lineare è lineare rispetto ai parametri α e β, anche se la relazione tra X e Y non deve necessariamente essere lineare. Nei modelli lineari generalizzati (GLM), la variabile dipendente Y è modellata seguendo una distribuzione appartenente alla famiglia esponenziale. Una distribuzione è parte della famiglia esponenziale se la sua funzione di densità di probabilità (o funzione di massa di probabilità, nel caso discreto) può essere espressa nella forma:

$$f_Y(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

dove:

- θ è il parametro naturale della distribuzione,
- ϕ è il parametro di dispersione,
- $b(\theta)$ è una funzione che determina la forma della distribuzione,
- $c(y,\phi)$ è una funzione che non dipende da θ.

Questa forma generale include molte distribuzioni comuni, come:

1. **Distribuzione Normale**:
   - $Y \sim \mathcal{N}(\mu, \sigma^2)$
   - Forma esponenziale: $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2}$, $c(y,\phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$.
2. **Distribuzione Binomiale**:
   - $Y \sim \text{Binomiale}(n,p)$
   - Forma esponenziale: $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$, $b(\theta) = n\log(1+e^\theta)$, $c(y,\phi) = \log\left(\binom{n}{y}\right)$.
3. **Distribuzione Poisson**:
   - $Y \sim \text{Poisson}(\lambda)$

- Forma esponenziale: $\theta = \log(\lambda)$, $\phi = 1$, $b(\theta) = e^{\theta}$, $c(y, \phi) = -\log(y!)$.

Queste distribuzioni permettono di modellare variabili dipendenti Y che non sono normalmente distribuite, ampliando le applicazioni dei GLM rispetto ai modelli di regressione lineare tradizionali. La scelta della distribuzione appropriata dipende dalla natura dei dati e dal tipo di variabile dipendente che si sta modellando.

Nei modelli lineari generalizzati (GLM), la funzione di collegamento (*link function*) $g(\cdot)$ stabilisce una relazione tra il valore atteso della variabile dipendente Y, denotato come $E[Y|X]$, e una combinazione lineare delle variabili indipendenti. Questa relazione è espressa come:

$$g(E[Y|X]) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

dove η è il predittore lineare, α è l'intercetta, e $\beta_1, \beta_2, \ldots, \beta_p$ sono i coefficienti di regressione associati alle variabili indipendenti $X_1, X_2, \ldots, X_p$.

La funzione di collegamento $g(\cdot)$ permette di modellare relazioni non lineari tra X e Y, trasformando il valore atteso di Y in modo che possa essere espresso come una combinazione lineare delle variabili indipendenti. Ad esempio, nel caso di una regressione logistica, la funzione di collegamento è il logit, definito come:

$$g(E[Y|X]) = \log\left(\frac{E[Y|X]}{1 - E[Y|X]}\right)$$

Questa trasformazione consente di modellare la probabilità che Y assuma un certo valore in funzione delle variabili indipendenti, pur mantenendo la linearità nei parametri.(James et al. 2013)

3.2 Modello Esplicativo

L'analisi esplicativa utilizza, per ciascun donatore, il conteggio totale di donazioni accumulate nel periodo 2009–2023 come variabile risposta discreta $Y_i = \text{total\_donations}_i$. Le covariate includono età (all'ultimo anno osservato), classi d'età, anno della prima donazione e genere. Per limitare l'influenza di outlier rari si esclude la coda estrema ($Y_i < 100$), coerentemente con i vincoli clinici annui e con la cadenza osservativa.

3.2.1 Quasi-Poisson

Teoria del quasi-Poisson

La teoria del quasi-Poisson è un'estensione del modello di regressione di Poisson utilizzata nei modelli lineari generalizzati (GLM) per gestire dati di conteggio che

mostrano overdispersione. L'overdispersione si verifica quando la varianza dei dati è maggiore della media, una situazione che il modello di Poisson standard non può gestire poiché assume che la varianza sia uguale alla media.

Nel modello di Poisson standard, la distribuzione di Y è definita come:

$$Y \sim \text{Poisson}(\lambda)$$

dove λ è il parametro di intensità, e la varianza è uguale alla media: $Var(Y) = E[Y] = \lambda$.

Nel modello quasi-Poisson, invece, la varianza è proporzionale alla media, ma con un fattore di dispersione ϕ:

$$Var(Y) = \phi \cdot \lambda$$

dove ϕ è il parametro di dispersione che permette di modellare l'overdispersione. Quando $\phi > 1$, indica che c'è overdispersione nei dati.

La funzione di collegamento nel modello quasi-Poisson è la stessa del modello di Poisson, tipicamente il logaritmo naturale:

$$g(E[Y|X]) = \log(\lambda) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Il modello quasi-Poisson utilizza la stessa struttura lineare per il predittore η, ma permette una varianza maggiore rispetto alla media, fornendo una maggiore flessibilità nel modellare dati di conteggio che non seguono strettamente la distribuzione di Poisson.

Risultati

Nel nostro dataset la dispersione stimata risulta nettamente maggiore di 1, evidenziando overdispersione marcata rispetto al Poisson puro. A parità di link (log) i coefficienti mostrano pattern interpretabili: un effetto medio decrescente dell'età (attenuato nelle classi centrali quando si usa la categorizzazione), e un incremento significativo per i donatori abituali rispetto ai non periodici. Le incertezze correttamente "allargate" da $\hat{\phi}$ evitano eccessi di significatività dovuti alla varianza sottostimata dal Poisson.

3.2.2 Tweedie *(power $\sim$ 1.19)*

Teoria della Tweedie

Il modello Tweedie è un tipo di modello lineare generalizzato (GLM) che gestisce dati che possono avere una distribuzione di probabilità con una combinazione di caratteristiche di distribuzioni di Poisson e gamma. È particolarmente utile per modellare dati che includono valori zero e continui positivi, come i dati di assicurazione che comprendono sinistri con importi variabili.

Caratteristiche del Modello Tweedie

Il modello Tweedie appartiene alla famiglia esponenziale e si caratterizza per avere una funzione di varianza della forma:

$$Var(Y) = \phi \cdot \mu^p$$

dove:

- $\mu = E[Y]$ è il valore atteso,
- ϕ è il parametro di dispersione,
- p è il parametro di potenza che determina la forma della distribuzione.

Differenze rispetto al Quasi-Poisson

Il modello quasi-Poisson, come discusso in precedenza, assume che la varianza sia proporzionale alla media ($Var(Y) = \phi \cdot \lambda$), il che è utile per gestire l'overdispersione nei dati di conteggio.

Il modello Tweedie, invece, generalizza ulteriormente questa relazione introducendo il parametro di potenza p, che permette di modellare una gamma più ampia di distribuzioni:

- $p = 1$: corrisponde al modello di Poisson, dove la varianza è uguale alla media.
- $p = 2$: corrisponde al modello gamma, utilizzato per dati continui positivi.
- $1 < p < 2$: rappresenta una distribuzione Tweedie, che combina caratteristiche di Poisson e gamma, utile per dati con valori zero e continui positivi. Verrà utilizzata successivamente.

Funzione di Collegamento

Come nei modelli GLM, il modello Tweedie utilizza una funzione di collegamento per stabilire la relazione tra il valore atteso e una combinazione lineare delle variabili indipendenti:

$$g(E[Y|X]) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

La scelta della funzione di collegamento dipende dalla natura dei dati e dal valore del parametro di potenza p.

Risultati

Il tuning del parametro di potenza suggerisce $p \approx 1.19$, collocando la risposta tra Poisson e Gamma. A parità di specifica, la devianza/AIC risultano favorevoli al Tweedie rispetto al quasi-Poisson, segno che la legge di varianza μ^p cattura meglio l'eteroschedasticità e la struttura della coda. I coefficienti mantengono interpretazione su scala log-tasso, con pattern coerenti a quelli osservati nel quasi-Poisson.

3.2.3 Gamma

Teoria della distribuzione Gamma

Il modello gamma è utilizzato per modellare variabili dipendenti che sono continue e positive. È particolarmente utile per dati che rappresentano tempi di attesa, costi, o altre misure che non possono assumere valori negativi.

Caratteristiche del Modello Gamma

La distribuzione gamma è parte della famiglia esponenziale e ha una funzione di densità di probabilità definita come:

$$f_Y(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

dove:

- α è il parametro di forma,
- β è il parametro di scala,
- $\Gamma(\alpha)$ è la funzione gamma.

Nel contesto dei GLM, la varianza della distribuzione gamma è proporzionale al quadrato della media:

$$Var(Y) = \phi \cdot \mu^2$$

dove $\mu = E[Y]$ è il valore atteso e ϕ è il parametro di dispersione.

Funzione di Collegamento

Il modello gamma utilizza una funzione di collegamento per stabilire la relazione tra il valore atteso e una combinazione lineare delle variabili indipendenti. Una scelta comune per la funzione di collegamento nel modello gamma è il logaritmo naturale:

$$g(E[Y|X]) = \log(\mu) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Questa funzione di collegamento è appropriata perché garantisce che il valore atteso μ sia sempre positivo, riflettendo la natura dei dati modellati.

Tuttavia la sua funzione di collegamento canonica sarebbe la funzione inversa.

Risultati

La distribuzione Gamma, essendo una distribuzione continua e positiva, non è teoricamente ideale per modellare le donazioni di sangue, che sono dati discreti e limitati nell'intervallo $[0, 4]$. In teoria, una distribuzione discreta, come la binomiale, potrebbe essere più appropriata per rappresentare questo tipo di dati, ma nella sua applicazione ha dato scarsi risultati.

Infatti, la distribuzione Gamma offre una notevole flessibilità, a differenza della binomiale, grazie ai suoi iperparametri, che permettono di adattare la forma della distribuzione alle caratteristiche specifiche dei dati. Questa capacità di adattamento può risultare vantaggiosa in pratica, consentendo di ottenere un buon fit dei dati osservati, anche se la distribuzione non corrisponde perfettamente alla natura discreta delle donazioni di sangue.

Mentre la scelta della distribuzione Gamma potrebbe non essere teoricamente perfetta per dati discreti e limitati, la sua capacità di adattarsi bene ai dati grazie alla flessibilità dei suoi iperparametri la rende una scelta pratica in molti contesti.

3.3 Modello Predittivo

La densità Gamma in parametrizzazione shape–rate è

$$f_Y(y \mid \alpha, \beta) \;=\; \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, y^{\alpha-1} e^{-\beta y}, \qquad y > 0,$$

dove α è il parametro di forma e β il parametro di tasso (rate). In parametrizzazione shape–scale $\theta = 1/\beta$ si ha $f(y) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} y^{\alpha-1} \exp(-y/\theta)$.

Nel GLM Gamma vale $\mathrm{Var}(Y) = \phi\, \mu^2$ e il link canonico è l'inverso $g(\mu) = 1/\mu$ (il link log è spesso preferito per interpretabilità moltiplicativa e positività di μ).

3.3.1 Effetto Covid

Si introduce una dummy COVID pari a 1 per gli anni 2020–2022 e 0 altrimenti, con possibile interazione con l'età per cogliere effetti eterogenei:

$$\log \mu_{i,t} \;=\; \alpha + \beta_1\, y_{i,t-1} + \beta_2\, y_{i,t-2} + \gamma_1\, \mathrm{gender}_i + \gamma_2\, \mathrm{age}_{i,t} + \delta\, \mathrm{COVID}_t + \delta_{\mathrm{int}}\, \mathrm{age}_{i,t} \times \mathrm{COVID}_t.$$

Nei dati si osserva una contrazione media del tasso di donazione durante il triennio 2020–2022, con attenuazione per fasce d'età più elevate (interazione positiva).

3.3.2 Modello finale

Per la previsione one-step-ahead si impiega un GLM con link log; due alternative pratiche:

- quasi-Poisson (robusto a over/underdispersione):

$$Y_{i,2023} \;\sim\; \mathrm{QP}(\mu_{i,2023}, \phi), \qquad \log \mu_{i,2023} = x_i^{\top} \beta,$$

- Tweedie con p calibrato (qui $p \approx 1.2$):

$$Y_{i,2023} \;\sim\; \mathrm{Tw}(\mu_{i,2023}, \phi, p), \qquad \log \mu_{i,2023} = x_i^{\top} \beta.$$

Il vettore x_i include lag storici (almeno $y_{i,2022}, y_{i,2021}$), genere, età (o classi), indicatore COVID e, se utile, l'anno della prima donazione per catturare la seniority. La selezione del modello si basa su devianza/AIC e performance su test; il Tweedie tende a prevalere quando la distribuzione dei conteggi mostra molti zeri e varianza non lineare in μ.

4 Bayesian Hidden Markov Models

4.1 Accenni di Teoria

4.1.1 Processo Markoviano

Un processo stocastico è una collezione di variabili aleatorie indicizzate da un parametro, tipicamente il tempo, che può essere continuo o discreto ma con la condizione che sia equidistante. La sequenza si denota come $\{X_t\}_{t\geq 1}$. Le variabili aleatorie possono essere continue o discrete e possono essere correlate tra loro. Questa dipendenza caratterrizza alcune delle proprietà che il processo può assumere come la stazionarietà, l'indipendenza o la correlazione. Queste proprietà possono definire categorie di processi stocastici, ad esempio i processi di Poisson sono caratterizzati da eventi che si verificano in modo indipendente e i processi stazionari spmp caratteroszati da proprietà statiche, che non cambiano nel tempo.

Un processo stocastico è una collezione di variabili aleatorie indicizzate da un parametro (tipicamente il tempo). Nel caso discreto a passi equispaziati, la sequenza si denota come $\{X_t\}_{t\geq 1}$. Un processo è markoviano di ordine 1 se il futuro dipende dal presente ma non dal passato più remoto.

Il processo markoviano è un caso specifico dei processi stocastici, e in questo caso la proprietà principale è che lo stato futuro dipende solamente dal suo presente e non dal passato. L'ipotesi assuntiva è molto forte e permette di semplificare drasticamente un modello di serie storiche. Infatti non servirà più osservare l'intero passato per predire il futuro, ma basterà appena avere le informazione sullo stato precedente. Ciò si basa su proprietà base della probabilità, nel caso di un processo markoviano di primo ordine, l'osservazione successiva dipenderà solamente dalla precedente:

Definizione 4.1 (Proprietà di Markov)**.** Sia $\{X_t\}$ un processo stocastico. Allora $\{X_t\}$ è markoviano di ordine 1 se e solo se

$$\Pr(X_t \mid X_{t-1}, \ldots, X_1) \;=\; \Pr(X_t \mid X_{t-1}) \qquad t \geq 2.$$

Per la Definizione 4.1, tutte le future donazioni sono indipendenti da quelle passate, ad eccezione di k osservazion più recenti, dove k è il grado della catena markoviana: $x_{n+1} \perp x_{n-1} | x_n$.

Una catena di Markov omogenea (stazionaria nel tempo) è caratterizzata da:

- la matrice di transizione $\mathbf{A}$ con elementi $a_{ij} = \Pr(X_t = j \mid X_{t-1} = i)$, per $t \geq 2$;
- il vettore di probabilità iniziali $\boldsymbol{\pi}$ con $\pi_i = \Pr(X_1 = i)$;
- lo spazio degli stati finiti $Z = \{1, \ldots, K\}$.

Proprietà strutturali

Di seguito alcune proprietà importanti che verrano usate in seguito per la descrizione e definizione degli Hidden Markov Model in generale e di quello specificatamente utilizzato nel progetto.

Definizione 4.2 (Ergodicità (catene omogenee))**.** Sia $\{X_t\}_{t\geq 0}$ una catena di Markov a tempo discreto su spazio degli stati finito Z con matrice di transizione $\mathbf{A}$ (costante nel tempo). Se la catena è irriducibile, aperiodica e positivamente ricorrente, allora esiste ed è unica una distribuzione stazionaria $\boldsymbol{\pi}^*$ tale che $\boldsymbol{\pi}^*\mathbf{A} = \boldsymbol{\pi}^*$. Levin, Peres, e Wilmer (2009).

Il tempo di mixing misura quanto rapidamente la distribuzione della catena "dimentica" le condizioni iniziali e si avvicina a $\boldsymbol{\pi}^*$.

Definizione 4.3 (Catena di Markov omogenea nel tempo)**.** Una catena di Markov discreta $\{X_t\}_{t\geq 0}$ su spazio degli stati finito Z si dice omogenea (o stazionaria nel tempo) se esiste una matrice di transizione costante $\mathbf{A}$ tale che, per ogni $t \geq 1$ e per ogni $i, j \in Z$,

$$\Pr(X_t = j \mid X_{t-1} = i) = A_{ij},$$

ossia le probabilità di transizione non dipendono dal tempo. Una distribuzione $\boldsymbol{\pi}$ si dice stazionaria se soddisfa

$$\boldsymbol{\pi}\,\mathbf{A} = \boldsymbol{\pi}, \qquad \boldsymbol{\pi}\,\mathbf{1} = 1, \quad \boldsymbol{\pi} \geq \mathbf{0}.$$

Una catena, invece, è detta non omogenea se la transizione al tempo t è data da una matrice $\mathbf{A}_t$ che può dipendere da t o da covariate x_t, cioè $\mathbf{A}_t = \mathbf{A}(x_t)$. In generale:

- non esiste una distribuzione stazionaria unica (indipendente dal tempo) che soddisfi $\boldsymbol{\pi}\,\mathbf{A}_t = \boldsymbol{\pi}$ per tutti i t;
- si usa la nozione di "ergodicità debole": le distribuzioni indotte da condizioni iniziali diverse si avvicinano tra loro sotto opportune condizioni di contrazione.

4.1.2 Hidden Markov Models (HMM)

Come citato nel capitolo precedente, le catene markoviano sono molto utili perché riescono a riassumere tutta la storia passata con appena k passi precedenti, dove k è l'ordine della catena markoviana. Quest'ipotesi permette di ridurre la fattorizzazione passando da

$$Pr(x_1, \dots, X_N) = Pr(x_1)Pr(x_2|X_1)Pr(x_3|x_2, X_1) \dots Pr(x_n + 1|x_n, \dots, x_{n-k+1}) \dots$$

In questo modo il modello è computazionalmente efficiente in quanto ha bisogno di processare molta meno informazione.

Tuttavia, ciò non permette di modellare strutture più complesse, e in questo caso bisogna raggirare la proprietà di Markov (Definizione 4.1). Si introducono quindi delle variabili latenti, delle variabili nascoste (*hidden*), che non sono osservate. Vengono costituiti due stage/serie dipendenti tra loro. La prima serie è costituita da una catena markoviana di variabili latenti, Z_t al tempo t, che non possiamo osservare e che dipendono solo dallo stato precedente, Z_{t-1} allo stato $t-1$. La seconda serie, invece, sono le variabili osservate, e la loro emissione (distribuzione) dipende dallo stato latente.

Con parametri $\theta = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi})$ la fattorizzazione congiunta è:

$$\Pr(\mathbf{y}, \mathbf{z} \mid \theta) = \Pr(z_1 \mid \boldsymbol{\pi}) \prod_{t=2}^{T} \Pr(z_t \mid z_{t-1}, \mathbf{A}) \prod_{t=1}^{T} \Pr(y_t \mid z_t, \boldsymbol{\phi}).$$

In aggiunta gli HMM permettono di modellare non solo una sequenza ma anche molteplici.

Estensioni degli HMM

Gli Hidden Markov Model possono essere estesi a problemi più generalizzati. Si possono quindi rompere o adattare alcune parti della sua struttura, come la matrice di transizione e le probabilità di emissione. Facendo ciò si aumenta il grado di complessità del modello perdendo leggermente la sua facile intuitività. Ma allo stesso tempo si vanno a migliorare notevolmente le sue prestazioni e la sua adattabilità al mondo reale. Di seguito vengono riportate alcune delle estensioni che verranno poi implementato nel modello predittivo.

Emissioni flessibili

Secondo Fan (2015), dato un HMM del primo ordine e che sia almeno ergodico (Definizione 4.2), allora si può definire una famiglia esponenziale per la distribuzione condizionata alle osservazioni, ovvero la distribuzione delle emissioni. Normalmente le emissioni sono dei parametri della distribuzione di Poisson, λ, e della distribuzione Gaussiana, μ o σ. In questo caso si andrebbe ad estendere il problema e si ipotizza una

famiglia esponenziale (vedi Sezione 3.1.4), che racchiude sia la distribuzione Poisson che la Gaussiana. I parametri della famiglia esponenziale, o della distribuzione scelta, non dipenderebbero unicamente dallo stato latente, ma, dipenderebbero da dei coefficienti β_k e da delle covariate W_t. Le covariate sono note, e possono cambiare nel tempo, permettendo al modello una flessibilità temporale garantendo la stazionarietà, nel caso in non venisse implementato un sistema di covariate come si vedrà nella sezione successiva. I parametri da stimare diventano, quindi, i β_k, e il loro valore sarà condizionato allo stato latente. Di conseguenza, si produrrano K GLM dove K è il numero di stati latenti.

Non omogeneità

Per aumentare le performance del modello o per studiare a fondo le dinamiche che possono influire nel cambio di stato, allora può essere interessante aggiungere delle covariate nella matrice di transizione. Il processo è simile a quello descritto nella sezione precedente (vedi Sezione 4.1.2), tuttavia in questo caso non garantiremmo più la proprietà di omogeneità (vedi Definizione 4.3).

4.1.3 Bayesiana

4.2 Il modello applicato alle donazioni

4.2.1 Pyro

Dopo la parte teorica, in questo capitolo parleremo della applicazione della teoria ai dati sulle donazioni di sangue. Se nel precedentemente capitolo era stato utilizzato quasi totalmente il linguaggio di programmazione statistica R, e i pacchetti `tidyverse` e `tidymodels`, in questo capitolo i modelli verranno sviluppati attraverso il linguaggio di programmazione python e il pacchetto `Pyro`.

Pyro è un linguaggio universale di programmazione probabilistica scritto in Python e costruito con `PyTorch` nel *backend*. Ma a cosa serve? Secondo Bingham et al. (2018), Pyro permette di scrivere in modo facile e flessibile modelli probablistici, tra cui modelli di *deep learning* e bayesiani. I punti di forza di Pyro sono i seguenti:

- universale, in quanto può rappresentare qualsiasi distribuzione di probabilità computabile;
- scalabile, ideale per lavorare su grandi *data set*;
- minimale, è costituito da un piccolo ma potente *core*, e
- flexible, è facilmente automatizzabile, ma al contempo controllabile, quando necessario.

Ovviamente non è l'unico strumento utilizzabile per costruire i modelli mostrati successivamente. Una valida alternativa è `stan` e implementata in R attraverso i pacchetti come `rstan`.

4.2.2 Dati e covariate

I dati sono stati inizialmente elaborati in R, succesivamente esportati in formato tabellare (CSV), ed infine importati in Python. L'obiettivo era quello di ottenere dei dati in formato tabulare dove ogni riga rappresentasse il donatore e con una colonna per ogni anno d'osservazione, contenente il numero di donazioni effettuate in quello specifico anno. Infine, altre colonne con le informazioni sul donatore, come l'anno di nascita, il genere e l'età. Successivamente, i dati sono stati modificati ulteriormente addattandoli all'ambiente di sviluppo in Python.

Abbiamo ottenuto così un un pannelo longitudinale con $N = 9236$ donatori e le loro donazioni negli anni 2009-2023 ($T = 15$) e le covariate con le informazioni demografiche sul donatore. Per la modellizzazione dei dati si è diviso il dataset iniziale in 4 dataset più piccoli:

1. il primo contenente le colonne costituite dai conteggi annuali per donatore;
2. il secondo contenente le covariate necessarie per la determinazione delle probabilità iniziali;
3. il terzo costituito dalle covariate per la matrice di transizione,
4. e l'ultimo contenente le covariate per i modelli GLM.

Dataset sulle osservazioni delle donazioni

Il primo dataset sarà costituito da sole $y_t \in \{0, 1, 2, 3, 4\}$, dove t è il tempo. Il limite superiore del dominio dei dati è dovuto a una regola clinica: ovvero un donatore può al più donare 4 volte all'anno per sangue intero. Questo ha portato all'utilizzo di distribuzioni probabilistiche improrie, ovvero che non erano contemplate per la natura dei dati, ma che a causa della loro forma, si adattavano molto bene al caso nostro.

La distribuzione di probabilità più appropriata sarebbe stata infatti una distribuzione binomiale, o una categorica. Tuttavia la distribuzione di Poisson era quella che meglio si adattava i dati, anche se è stata troncata sulla coda destra a causa della regola clinica.

Questi dati non hanno avuto bisogno di attenzione, in quanto sono le osservazioni dei donatori e l'output del nostro modello.

Dataset con le covariate per le probabilità iniziali

Il secondo dataset sarà usato per calcolare $\pi = Pr(Z_{n,1} \mid X^{\pi}_{n,1})$, ovvero la probabilità dell'individuo n di iniziare nello stato k, date le covariate $X^{\pi}_{n,1}$, dove $X^{\pi}_{n,1}$ è una matrice di dimensione N, P_{π} con p_{π} il numero di covariate utilizzate.

Le probabilità iniziali vengono calcolate solo per il primo stato. Devono, perciò, sfruttare delle informazioni disponibili al tempo 1. Le sue covariate saranno di conseguenza statiche.

La sua importanza è relativa, infatti grazie alla matrice di transizione, le unità si sposteranno tra gli stati e assumeranno lo stato che più gli si adatta allo specifico tempo t.

Dataset con le covariate per la matrice di transizione

Il modello che verrà sviluppato in questo capitolo è diverso da un classico *hidden markov model* e non possiede tutte le sue proprietà. In specifico non gode della proprietà di omogeneità, ovvero non è un modello stazionario (vedi Definizione 4.3). Non ci aspettiamo che un donatore appartenga sempre allo stato, ovvero che sia propenso sempre allo stesso modo di donare. Riteniamo che esso dipenda da fattori esterni che influenzino il suo comportamento. Questi fattori non sono fissi e cambiano nel tempo. Sarà dunque necessario includere delle covariate dinamiche, che cambiano nel tempo, come l'età del donatore. Otterremo dunque una matrice di 3 dimensioni N, P_A, T dove N è il numero di donatori, P_A il numero di variabili usate e T gli anni usati dal modello.

Le covariate utilizzate saranno l'età e gli anni del covid. Analizzando l'età, si osserva come il suo andamento non sia lineare. Al crescere dell'età non si può assertare né che il numero di donazioni sia in aumento, né che il numero di donazioni sia in diminuzione. Infatti agli estremi, ovvero nelle fasce d'età più giovani e più anziane si osserveranno un numero di donazioni minore mentre nella fascia centrale, un numero di donazioni maggiore. Sia $\mathbf{X}$ e $\mathbf{Y}$ due vettori (o matrici) e sia $\mathbf{A}$ una matrice di trasformazione tale che

$$\mathbf{Y} = \mathbf{A}\,\mathbf{X} + \boldsymbol{\varepsilon}.$$

Il nostro obiettivo è trovare

$$\arg \min_{\mathbf{A} \in \mathbb{R}^{q \times p}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|^2.$$

Le opzioni possibili per gestire il suddetto problema sono molteplici, dalle più semplici alle più complesse.
Tra le principali strategie si possono citare:

- **aggiunta di potenze della variabile**: ad esempio includendo termini quadratici o cubici dell'età ($\text{età}^2, \text{età}^3, \ldots$) per catturare eventuali relazioni non lineari;

- **categorizzazione della variabile**: suddividere l'età in fasce (es. 18–24, 25–34, 35–44, ...) e introdurre variabili dummy per ciascuna categoria, consentendo al modello di apprendere effetti differenti per ciascun gruppo;
- **funzioni spline**: utilizzare spline lineari o cubic spline per rappresentare in modo flessibile l'andamento non lineare dell'età, mantenendo al contempo una continuità nei punti di raccordo;
- **altri approcci di regolarizzazione o riduzione di dimensionalità**: ad esempio penalizzazioni (ridge, lasso) per evitare overfitting in presenza di molte trasformazioni, oppure tecniche di smooth fitting per controllare la complessità del modello.

Queste trasformazioni permettono di modellare meglio l'effetto dell'età, evitando di imporre una relazione lineare troppo restrittiva e garantendo al tempo stesso una maggiore flessibilità nella costruzione della matrice di transizione.

Tra le diverse strategie illustrate, in questo lavoro si è deciso di adottare la **categorizzazione della variabile età**.
Questa scelta consente di distinguere in modo chiaro le diverse fasce d'età dei donatori, mantenendo al tempo stesso un modello interpretabile. In particolare, le categorie permettono di osservare come il comportamento donazionale vari nei diversi gruppi, evidenziando ad esempio differenze tra i donatori più giovani, quelli in età intermedia e i più anziani.
In questo modo si ottiene un compromesso tra semplicità e capacità di rappresentare l'andamento non lineare della variabile, evitando di introdurre complessità eccessiva tramite spline o polinomi di grado elevato.

Durante la pandemia da coronavirus SARS-CoV-2 la propensione a donare e le sue motivazioni sono cambiate molto. Riteniamo quindi di dovere aggiungere quest'informazione al modello e manipolare la serie storica in questo momento storico in modo a se stante. Infatti grazie all'introduzione di questa covariate si possono osservare in quegli anni, un profondo cambiamento del trend.

Dataset con le covariate per i GLM

Attraverso le probabilità iniziali e la matrice di transizione possiamo risalire allo stato latente dell'individuo lungo la sua serie storica. Con l'obbietivo di predire il valore atteso della prossima donazione, lo stato latente più importante che ci interesserà conoscere, sarà l'ultimo stato latente, ovvero lo stato che ci dirà quale dei k dei K diversi modelli utilizzare. Il HMM-GLM ha esattamente questo di differenza confronto un normale HMM, l'HMM-GLM non restituisce un *rate*, o il parametro di una qualsiasi distribuzione; ma restituisce l'indicazione a quale modello usare. Infatti, ipotizziamo che il modello generatore dei dati sia significativamente diverso tra i 3 diversi *pattern* di dontari. E non è una differenza fissa/statica, ma una differenza molto più complessa, che è influenzata anche da altri fattori. Per questo, si ritene opportuno usare 3 modelli diversi per modellare e predire le future donazioni di sangue.

Le covariate scelte in questo caso sono un incrocio tra covariate statiche e dinamiche. Le variabili scelte sono il genere, le fasce d'età e il periodo covid.

Dopodiché si è diviso il dataset iniziale in 4 dataset più piccoli.
- Covariate disponibili: genere (M/F), anno di nascita. Covariate costruite:

- Iniziali $x^{\pi} \in \mathbb{R}^3$: intercetta, birth_year_norm (z-score), gender_code (F=1, M=0).
- Transizioni $x_t^A \in \mathbb{R}^8$: intercetta; fasce d'età one-hot (baseline 18-24, poi 25-34, 35-44, 45-54, 55-59, 60-64, 65+); indicatore COVID (1 per 2020-2022, 0 altrimenti).
- Emissioni $x_t^{\text{em}} \in \mathbb{R}^9$: intercetta; gender_code; stesse dummies di età; COVID.

[**Da integrare: specifica finale e grafico delle fasce d'età.**]

4.2.3 Componenti del Modello

Il modello, come citato in Sezione 4.2.1, è definito tramite Pyro. Si ritiene che sia importante esplicitare il codice, tanto quanto la stesura di una formula matematica. Mediante la lettura del codice si comprende a pieno le potenzialità di Pyro e la struttura del modello.

Il numero di stati latenti, `K`, è stato fissato a 3. Quindi, $z_{n,t} \in \{0, 1, 2\}$, e non sono osservate. La loro interpretazione reale sarà il profilo di donatore, ovvero se sarà un donatore frequente, occasionale o sporadico.

Le osservazioni, $y_{n,t} \in \mathbb{N}$, sono conservate nella variabile `obs`, che è una matrice di dimensione `N`, `T` e assume valori in $[0, 4]$. Come spiegato nel capitolo precedente (vedi Sezione 4.2.2), sono presenti altre tre matrici composte da:

- covariate sulle probabilità inziali, $W_\pi \in \mathbb{R}^{K \times 2}$;
- covariate sulla matrice di transizione, $W_A \in \mathbb{R}^{K \times K \times 8}$; e
- covariate sulle emissioni, $\beta_{em} \in \mathbb{R}^{K \times (1+9)}$, che in questo caso saranno le covariate dei K diversi modelli.

In Python ci sarà una prima parte di configurazione del modello, con le diverse dimensioni degli oggetti da passare, la scelta del numero di stati.

Approccio Bayesiano

Abbiamo adottato un approccio bayesiano per le componenti che governano le probabilità iniziali e le transizioni dell'HMM, usando prior di tipo Dirichlet sulle intercette e lasciando stimare per MAP (via `pyro.param`) le pendenze che modulano l'effetto delle covariate. In particolare:

- distribuzione iniziale: $\pi_{\text{base}} \sim \text{Dirichlet}(\boldsymbol{\alpha}_\pi)$ con $\boldsymbol{\alpha}_\pi$ asimmetrica per spezzare la simmetria delle etichette e incorporare la conoscenza a priori sull'elevata prevalenza dello "stato inattivo" (non-donatore), dovuto al fatto che al tempo $t = 0$, ancora molti dei donatori non avevano iniziato a donare;
- matrice di transizione: per ciascuna riga k, $A_{\text{base}}[k, \cdot] \sim \text{Dirichlet}(\boldsymbol{\alpha}_{A_k})$. Abbiamo valutato sia una a priori non informativa (simmetrica) sia una a priori "diagonale" (con massa concentrata sulla permanenza $k \to k$).

Motivazioni per l'approccio bayesiano:

- identificabilità pratica: una prior asimmetrica su π mitiga il *label switching* e rende più stabile il confronto tra fit riavviati con semi diversi;
- inferenza scalabile: l'uso della variazionale stocastica con enumerazione esatta delle variabili discrete (`TraceEnum_ELBO`) riduce la varianza del gradiente e consente di trattare in modo efficiente le catene latenti $z_{n,t}$ {Hoffman et al. (2013)};
- coerenza interpretativa: la combinazione "prior diagonale" su A e Poisson/Gamma sulle emissioni porta a tre profili comportamentali stabili (non-, leggero, frequente donatore), con transizioni credibili e facilmente comunicabili.

Definzione del modello

Il modello è caratterizzato da due cicli annidati tra loro, uno sul numero di osservazioni e uno sulla serie temporale. Questi cicli sono caratterizzati da un passo base e un passo iterativo. Potremmo definire l'algoritmo così come segue:

Passo base n:

1. campionamento delle a priori e conversione in base logaritmica per le variabili $\pi_{\text{base},k}$ e $A_{\text{base},kj}$
2. calcolo dei coeffienti tramite la matrice delle covariate per i parametri $W_{\pi,k}$, W_A, kj e β_{em}, k.

Passo iterativo n:

Passo base t:

3. Calcolo delle probabilità inziale dell'individuo„ di iniziare la sua traiettoria nello stato k, tramite la formula

$$\Pr(z_{n,0} = k \mid x_n^{\pi}) = \text{softmax}_k\Big(\log \pi_{\text{base},k} + W_{\pi,k} \cdot x_n^{\pi}\Big)$$

e campionamento dello stato da tali probabilità.

4. Calcolo del parametro $\log(\mu_{n,t}^{\text{em}})$ corrispondente al logaritmo del valore atteso di donare al tempo $t = 0$. Campionamente di y_0 dalla distribuzione di Poisson con il parametro appena calcolato.

Passo iterativo t:

5. Calcolo delle probabilità di transizione al tempo t sfruttando la matrice di transizione calcolata nel passo base di n e pesata per le informazioni che si hanno a disposizione per l'unità n al tempo t. Attraverso questa formula

$$\Pr(z_{n,t} = j \mid z_{n,t-1} = k,\ x_{n,t}^{A}) = \text{softmax}_j\Big(\log A_{\text{base},kj} + W_{A,kj} \cdot x_{n,t}^{A}\Big)$$

si campiona la probabilià che l'unità n occupi lo stato j al passo t sapendo che al passo $t-1$ occupava lo stato k.

6. Dato lo stato j, calcolato al punto precedente, campionamento del numero di donazioni al tempo t:

$$y_{n,t} \mid z_{n,t} = k \sim \text{Poisson}\Big(\exp\Big(X_{n,t}^{em}\beta_{\mathbf{em},\mathbf{k}}\Big)\Big)$$

```
K      = 3      # Number of latent states
C_pi   = cov_init.shape[1]      # Initial-state covariates (birth_year_norm, gender
C_A    = cov_tran.shape[2]      # Transition covariates (ages_norm, covid_years)
C_em   = cov_emission.shape[2]      # Emission covariates (gender, 7 age dummies, c

# Dirichlet prior
alpha_pi = torch.tensor([5., 2., 1.])
alpha_A  = torch.ones((K, K))


@config_enumerate
def model(obs, x_pi, x_A, x_em):
    N, T = obs.shape

    # 1) Priors/parameters for initial and transition distribution
    pi_base = pyro.sample("pi_base", dist.Dirichlet(alpha_pi))          # [K]
    A_base  = pyro.sample("A_base",  dist.Dirichlet(alpha_A).to_event(1)) # [K, K]
    log_pi_base = pi_base.log()
    log_A_base  = A_base.log()
```

```
    # 2) Slope coefficients for initial and transition covariates
    # and State-specific GLM emission coefficients (learned)
    W_pi = pyro.param("W_pi", torch.zeros(K, C_pi))              # [K, C_pi]
    W_A  = pyro.param("W_A", torch.zeros(K, K, C_A))             # [K, K, C_A]
    beta_em = pyro.param("beta_em", torch.zeros(K, C_em))        # [K, C_em+1]

    with pyro.plate("seqs", N):
        # Initial hidden state probabilities: depend on x_pi via W_pi
        logits0 = log_pi_base + (x_pi @ W_pi.T)                  # [N, K]
        z_prev = pyro.sample("z_0",
                             dist.Categorical(logits=logits0),
                             infer={"enumerate": "parallel"})

        # Emission at t=0: state-specific GLM on covariates x_em[:, 0, :]
        log_mu0 = (x_em[:, 0, :] * beta_em[z_prev, :]).sum(-1)
        pyro.sample("y_0", dist.Poisson(log_mu0.exp()), obs=obs[:, 0])

        # For t = 1 ... T-1, update state and emit
        for t in range(1, T):
            x_t = x_A[:, t, :]                                   # [N, C_A]
            logitsT = (log_A_base[z_prev] + (W_A[z_prev] * x_t[:, None, :]).sum(-1)
            z_t = pyro.sample(f"z_{t}",
                              dist.Categorical(logits=logitsT),
                              infer={"enumerate": "parallel"})

            # Emission: state-specific GLM at time t
            log_mu_t = (x_em[:, t, :] * beta_em[z_t, :]).sum(-1)
            pyro.sample(f"y_{t}", dist.Poisson(log_mu_t.exp()), obs=obs[:, t])
            z_prev = z_t
```

Definizione del guide

Il guide specifica la famiglia variazionale $q(\cdot)$ utilizzata da SVI per approssimare la posteriore. Qui scegliamo una approssimazione "a massa puntuale" (Delta) per i soli nodi continui latenti del modello, cioè le intercette della distribuzione iniziale e della matrice di transizione: $q(\pi_{\text{base}}) = \delta(\pi_{\text{base}} - \hat{\pi})$ e $q(A_{\text{base}}) = \delta(A_{\text{base}} - \hat{A})$. Le variabili discrete $z_{n,t}$ non compaiono nel guide perché vengono enumerate esattamente nel modello (`@config_enumerate`), e i coefficienti W_π, W_A, β_{em} sono stimati come parametri deterministici (`pyro.param`). In termini teorici, si massimizza l'ELBO scegliendo una famiglia variazionale degenere (MAP) per π_{base} e A_{base}, in modo da semplificare l'ottimizzazione e stabilizzare l'identificazione. Anche se, in questo modo, non si quantifica l'incertezza.

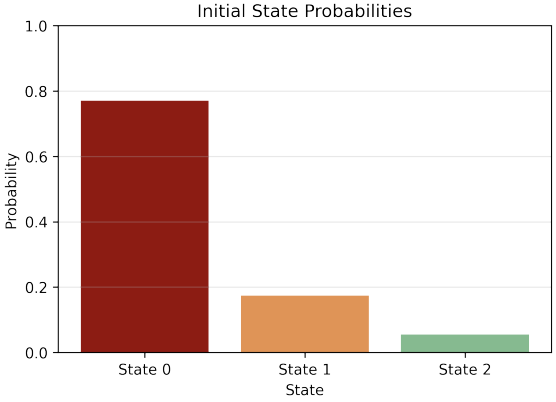
I vincoli simplex (`dist.constraints.simplex`) garantiscono che $\hat{\pi}$ e ciascuna riga di $\hat{A}$ siano probabilità valide (non negative e a somma unitaria). L'inizializzazione

deve già rispettare il vincolo: normalizziamo $\boldsymbol{\alpha}_\pi$ per ottenere un punto di partenza ammissibile per π_q, mentre per A_q usiamo una matrice "diagonale rinforzata" rinormalizzata per riga, che favorisce la persistenza.

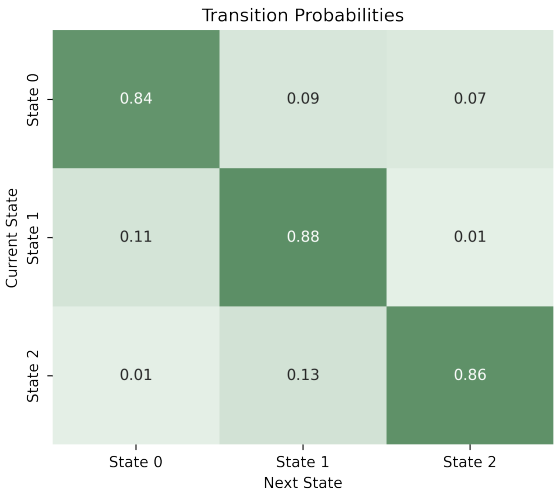
```
def guide(obs, x_pi, x_A, x_em):
    pi_q = pyro.param(
        "pi_base_map",
        alpha_pi,
        constraint=dist.constraints.simplex
    )

    # Each row: simplex for state-to-state transitions
    A_init = torch.eye(K) * (K - 1.) + 1.
    A_init = A_init / A_init.sum(-1, keepdim=True)
    A_q = pyro.param(
        "A_base_map",
        A_init,
        constraint=dist.constraints.simplex
    )

    pyro.sample("pi_base", dist.Delta(pi_q).to_event(1))
    pyro.sample("A_base",  dist.Delta(A_q).to_event(2))
```

Allenamento del modello

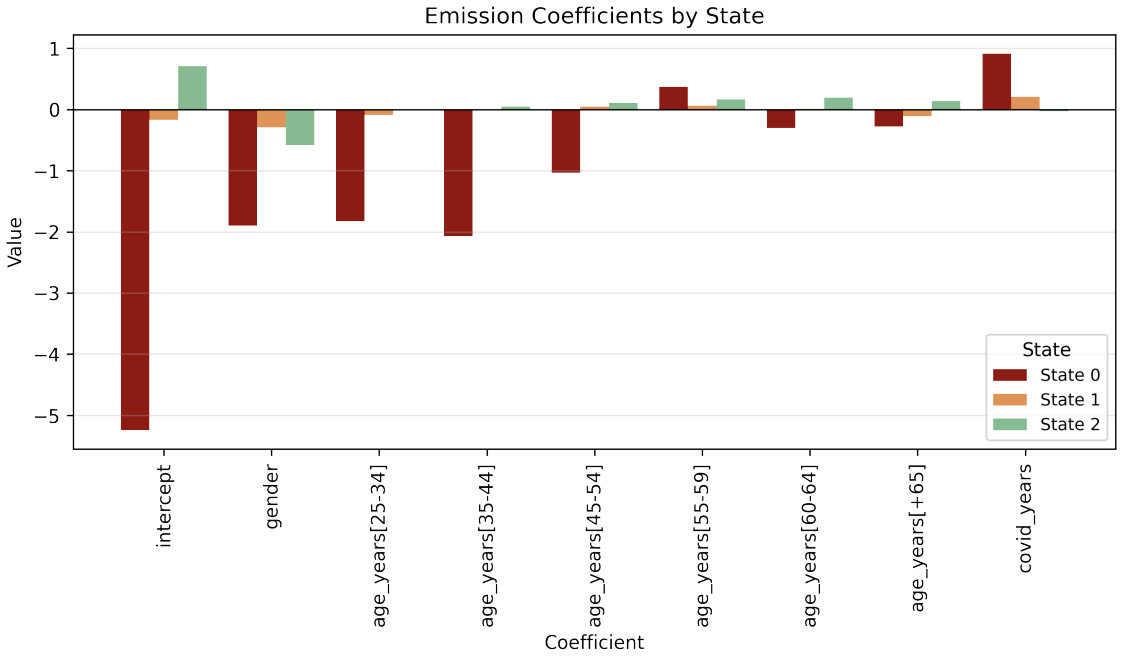
```
pyro.clear_param_store()
svi = SVI(model, guide,
          Adam({"lr": 2e-2}),
          loss=TraceEnum_ELBO(max_plate_nesting=1))

for step in range(2000):
    loss = svi.step(obs_torch, cov_init_torch, cov_tran_torch, cov_emiss_torch)
    if step % 200 == 0:
        print(f"{step:4d}  ELBO = {loss:,.0f}")

param_path = here("models/hmm_glm_full.pt")
pyro.get_param_store().save(param_path)
print(f"ParamStore saved in: {param_path}")
```



(a) Occupazione degli stati iniziali



(b) Matrice di transizione



(c) Coefficienti dei modelli GLM

Figura 4.1: Risultati principali del modello

4.3 Risultati

4.4 Algoritmi e diagnostiche

4.4.1 Viterbi con covariate

A parametri fissati (plug-in), il percorso MAP $z^*_{0:T}$ si ottiene per programmazione dinamica:

Inizializzazione

$$\delta_0(k) \;=\; \log \Pr(z_0{=}k \mid x^\pi) \;+\; \log \Pr(y_0 \mid z_0{=}k).$$

Ricorsione per $t = 1, \ldots, T$

$$\delta_t(j) \;=\; \max_k \Big\{ \delta_{t-1}(k) \;+\; \log \Pr(z_t{=}j \mid z_{t-1}{=}k, x^A_t) \Big\} \;+\; \log \Pr(y_t \mid z_t{=}j).$$

Backtracking: $z^*_T = \arg\max_k \delta_T(k)$, poi $z^*_{t-1} = \psi_t(z^*_t)$.

4.4.2 Forward per log-likelihood e previsione

Con la forward recursion si ottiene la log-likelihood e, condizionando su α_T, la previsione one-step-ahead:

Filtraggio

$$\alpha_t(j) \;\propto\; \Pr(y_t \mid z_t{=}j) \sum_k \alpha_{t-1}(k)\, \Pr(z_t{=}j \mid z_{t-1}{=}k, x^A_t).$$

Propagazione allo step successivo

$$p_{T+1}(j) \;=\; \sum_k \alpha_T(k)\, \Pr(z_{T+1}{=}j \mid z_T{=}k, x^A_{T+1}), \qquad \mathbb{E}[y_{T+1}] \;=\; \sum_{j=1}^{K} p_{T+1}(j)\, \lambda_{T+1}(j).$$

4.4.3 Occupancy e switch-rate

- Occupancy nel tempo: quota di donatori in ciascuno stato per anno; si osserva uno spostamento da stato 0 verso stati più attivi durante il 2020-2022 (indicatore COVID).
- Switch-rate “any switch”: circa 79.9% dei donatori cambia stato almeno una volta. Se si volesse un “tasso per step”, definire esplicitamente la metrica e ricalcolare.

[Da integrare: grafico di occupancy e misura quantitativa del delta pre/post 2020.]

4.4.4 Calibrazione e confronto con GLM

Valutazione one-step-ahead su test hold-out (90/10 stratificato):

- Mean log-likelihood per sequenza circa -13.05 (test), -13.47 (train);
- MAE circa 0.51; RMSE circa 0.77; Brier per $I\{y > 0\}$ circa 0.18; NLL circa 0.88 (test).

Confronto con GLM Poisson "vanilla":

- GLM: pred mean 0.96 (obs 0.91), MSE 1.016, accuracy(round) 28.2%;
- HMM mixture mean non filtrato: pred mean 0.55, MSE 1.248, accuracy(round) 40.0%.

Interpretazione: senza filtraggio, la mistura di Poisson è conservativa sul livello medio ma migliora l'accuracy su conteggi arrotondati; con filtraggio (uso di α_T) la calibrazione di $P(y > 0)$ è buona (Brier circa 0.18).

[Da integrare: reliability plot per $P(y > 0)$, PPC per distribuzioni annuali e per sottogruppi.]

4.5 Scelta del numero di stati

Criteri usati: trend di ELBO, log-likelihood su hold-out via forward, penalizzazione stile BIC

$$\log p(X) \;\approx\; \log p\Big(X \mid \hat{\theta}_{\mathrm{MAP}}\Big) \;-\; \tfrac{1}{2} M \log N.$$

Risultato: piccoli miglioramenti fino a $K \approx 4$; si adotta $K = 3$ per interpretabilità e stabilità.

[Da integrare: figura con log-evidence e BIC-like vs K e breve commento.]

4.6 Riproducibilità e note operative

- API operative: caricamento parametri, Viterbi, forward log-likelihood, metriche one-step, predizione per singolo donatore.
- Ambiente: riportare versioni PyTorch/Pyro, seed, path ai modelli (prod vs dev).
- Allineamento colonne: documentare l'ordine esatto di x^{em} e di $\boldsymbol{\beta}_{\mathrm{em}}$ per evitare mismatches.

Limite del numero di covariate al crescere del numero di stati.

5 Dashboard e Sito Web

5.1 Motivazione

Lo scopo della tesi non è sololo studio e l'analisi di un fenomeno da parte del candidato. La tesi in questione si propone come soluzione a un problema reale. Essa unisce il mondo accademico al mondo reale, il prodotto dev'essere qualcosa di tangibile e facilmente usufruibile. Se per il modello esplicativo è sufficiente esporre graficamente o tabularmente i risultatt, per il modello predittivo bisogna fare qualcosa di più. Come dice il nome stesso, lo scopo del modello predittivo è quello di calcolare il valore che ci si attende da una variabile date delle variabili di input. Il calcolo può essere svolto manualmente per un GLM, ma già per un modello come l'HMM-GLM visto nel capitolo precedente, il suo calcolo diventa notevolmente più complesso e l'utilizzo di un calcolatore risulta strettamente necessario. La dashboard permette quindi di

5.2 Sviluppo

- Front-end: **Quarto + Shiny**.
- Back-end: modelli salvati in formato `torch.script` / `.rds`, caricati on-demand.

6 Conclusioni

6.1 Risultati

6.2 Idee per il Futuro

- **Hierarchical HMM** per distinguere tipologia di emocomponenti (sangue intero, plasma, piastrine).

Riferimenti

Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, e Noah D. Goodman. 2018. «Pyro: Deep Universal Probabilistic Programming». *Journal of Machine Learning Research*.

Bordandini, Paola. 2025. «Geografie del capitale sociale. Trent'anni di senso civico in Italia». Riccardo Prandini. https://cris.unibo.it/retrieve/ee59ddb4-bdeb-4be5-8d2e-e70467252248/Donazioni%20di%20sangue.pdf.

Caselli, Graziella, Jacques Vallin, e Guillaume Wunsch. 2006. *Demografia. Analisi e sintesi*. Bologna: Il Mulino.

Fan, Jieyu. 2015. «On Markov and Hidden Markov Models with Applications to Trajectories».

Hoffman, Matthew D., David M. Blei, Chong Wang, e John Paisley. 2013. «Stochastic Variational Inference». *Journal of Machine Learning Research*.

ISTAT. 2023a. «Bilancio demografico e popolazione residente: definizioni e note metodologiche». Roma: Istituto Nazionale di Statistica.

———. 2023b. «Tavole di mortalità della popolazione residente: metodi e note tecniche». Roma: Istituto Nazionale di Statistica.

James, Gareth, Daniela Witten, Trevor Hastie, e Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer.

Levin, David A., Yuval Peres, e Elizabeth L. Wilmer. 2009. *Markov Chains and Mixing Times*. American Mathematical Society.

Pati, Ilaria, Claudio Velati, Carlo Mengoli, Massimo Franchini, Francesca Masiello, Giuseppe Marano, Eva Veropalumbo, et al. 2021. «A forecasting model to estimate the drop in blood supplies during the SARS-CoV-2 pandemic in Italy». *Transfusion Medicine* 31 (marzo): 200–205. https://doi.org/10.1111/tme.12764.