

ANALISI DELLA STORIA DEGLI EVENTI

Scopo: spiegare perché certi soggetti sono più favoriti a subire un evento rispetto ad altri. Ciò può essere ottenuto utilizzando particolari tipi di modelli che, a seconda dell'ambito in cui sono applicate, sono chiamati modello a tempo di fallimento, a tempo di vita, modello di sopravvivenza, modello sul tasso di decadenza, modello di pericoli.

Per es. il modello di reschio è un modello di regressione in cui il reschio si sperimenta un evento ad un momento di tempo certo e prevede da un insieme di covariate (variabili esplicative/dependenti)

Esempi:

- impatto della scolarizzazione sui tempi del matrimonio e della fertilità
- impatto delle differenze individuali nell'età al primo parto
- impatto del livello di istruzione nel caso morte sovrato a diverse cause
- impatto della mortalità professionale sui tempi del matrimonio

DISEGNO DELLO STUDIO

- a) dati di conteggio degli eventi: fix unità - o conta eventi registrano il numero dei vari tipi di eventi per ogni unità
- b) dati di sequenza degli eventi: fix unità - o conta flussi delle unità sequenze degli stati occupati da ciascuna unità
- c) dati trasversali incrociati: si prende una parte delle popolazioni, una sezione microscopica, e lì si studia per un certo tempo
 - 1) equilibrio statistico? si è entrata e si uscita da ciascuno stato deve essere regolare nel tempo
 - 2) indicazione di causa-effetto?
 - 3) fit di forze degli effetti reciproci?
 - 4) informatività?
 - 5) validità nelle dipendenze di stato? l'ingresso e l'uscita dagli stati sono altamente variabili nel tempo
 - 6) come i cambiamenti nelle variazioni esplicative generano variazioni nei risultati?

7) Date mancante?

d) dati panel: le unità sono seguite nel tempo

- 1) quadro d'errore: ≠ risposte a ≠ andate
- 2) modello del processo: gli studi panel influenzano i processi; è allo studio modifica e comportamento dei soggetti
- 3) logoramento del campione: campione diminuisce nel tempo
- 4) risposte e date mancanti
- 5) coatti fallaci: focus sulla coorte in periodo specifico
- 6) ritardo fra i tempi della causa e l'inizio dell'effetto
- 7) forme ≠ di come l'effetto si sviluppa nel tempo
- 8) difficile a catturare l'andamento di variabili dipendente dal tempo

e) dati basati sulla storia degli eventi: formano una misura continua di variabili qualsiasi e registrano variazioni di variabili qualsiasi e la loro temporistica

Di solito questi dati sono raccolti in modo retrospettivo attraverso studi sulla storia delle vite.

- 1) sono nuovi economici dei dati panel
- 2) sono codificati per un quadro specifico
- 3) problematici perché gli interventisti difficilmente ricordano la temporica delle avvertenze di alcun stato, in maniera accurata (sovrattutto in studi studiudinali e campionamenti)
- 4) non possono essere utilizzati per stabilire i fattori che includono variabili non note all'interventista
- 5) devono essere lasciati sui sopravvissuti (bias di selezione)

BIAS: è un errore sistematico presente in uno studio che si riferisce sui risultati determinando uno scarto fra risultati ottenuti e quelli che si sarebbero dovuti ottenere in assenza di bias.

3 bias più importanti sono:

- bias di selezione: si riferisce se il campione è scelto e assemblato in modo errato
- bias di misurazione: si riferisce se i metodi di misurazione sono infatti o non ben lavorati.
- bias da effetti estranei: è presente un fattore estraneo di confondimento

L'ideale sarebbe avere un disegno misto, fatto di follow-up e data che contengono

i loro punti di forza: i tradizionali disegni pieni con le vertutis dei dati degli eventi storici retrospettivi

I modelli di storia degli eventi sono un utile appoggio per scoprire relazioni causali.

ASSOCIAZIONE E NESSO DI CAUSALITÀ

La letteratura sociologica ricorda una sequenza di: complemente degli studi, matrimonio e nascita del primo figlio. Il discorso fra età al matrimonio e prima nascita è abbastanza stretto in Grecia (1,8 anni). Da questo stretto spazio fra i tempi del matrimonio e della nascita del primo figlio, e il fatto che la scolarizzazione e il matrimonio sono generalmente eventi incontrastabili, un accorgimento nella temporalità del matrimonio, come conseguenza di una maggiore scolarizzazione, si traduce direttamente in un cambiamento nel tempo di inizio della fertilità. D'altra parte, in un paese sviluppato, dove molte donne ritardano la nascita del primo figlio, un cambiamento nel tempo del matrimonio a causa della maggiore scolarizzazione non si traduce necessariamente in una ritardata fertilità.

- La scuola ha un impatto causale sui tempi del matrimonio in Grecia? La letteratura empirica ha generalmente confermato la connivenza negativa tra la scuola e l'età al momento del matrimonio. Ma cosa passa dietro l'impatto causale?

Esempio

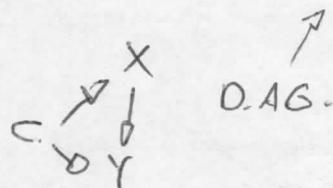
TH: valutare l'effetto di X su Y

$Y=1$ se sposati, $Y=0$ altrimenti

$X=1$ se laureati, $X=0$ altrimenti

C = religione (confondente)

grafico relazioni causali a priori
conosciute a conoscenze posteriori.



CONFONDENTE: è un fattore/effetto che confonde; distorce la capacità di attribuire la causa di qualcosa al trattamento, perché qualcosa altro potrebbe influenzare il risultato

es. se un gruppo di soggetti sta ricevendo più terapie, sarebbe possibile confondere l'effetto di una terapia con d'altra.

(relazione spuria) - Stratificando per una variabile, la var. di interest presenta valori uguali nei gruppi di stratificazione \neq dal suo valore grezzo (ma stava la differenza entro ame stratificate)

- va inserito nel modello (removal)

es. sole \rightarrow ipertensione
età

- deve: a) essere associato a X
- b) avere effetto su Y

c) non deve esser mediatore fra X e Y

(non deve trovarsi sul percorso fra X e Y)

Con la modellazione $E[Y|X] = g(x, \gamma)$ attraverso un modello logit siamo la associazione (γ) fra insiemismo e declarazione cioè vogliamo: $\Pr(Y=1|X=1) = \Pr(Y=1|X=0)$ cioè perciò vuol essere confuso da $C \rightarrow$ (altrimenti bisogna di controllare per poter stimare l'effetto causale fra X e Y).

Se le 2 probabilità sono =, non c'è associazione

L'associazione non implica la causalità, ma la causalità implica la associazione

L'associazione implica la causalità solo se non c'è confondimento che per le variabili misurate sia per le non misurate. Sia C il confondente: $Y = \alpha + \beta_1 X + \beta_2 C$ (aggiunto per C)

Il mediatore invece non va messo nel modello.

MEDIATORE: $X \rightarrow M \rightarrow Y$ (non lo vediamo)

nel modello

• se lo mettessi diminuirei l'effetto delle interazioni fra X e Y perché lo controllerei per M .

• tenendo conto dell'intervento di M l'associazione fra X e Y scompare

es. istruzione genitor -> estrazione figlio -> reddito figlio

• stratificando per una variabile i risultati sono molto \neq nei gruppi di stratificazione

OSS Spiegazioni alternative per una ostacolata associazione fra $X \rightarrow Y$:

• $T \stackrel{X}{\not\rightarrow} Y$ relazione spuria = CONFONDIMENTO

• relazione indipendente dai valori di un'altra variabile

• relazione dovuta alla presenza di un'altra variabile che interviene fra X e Y $X \rightarrow Y$

• errore campionario

VARIANZA
TERRITORIO
LIVELLI DI TUTTO

ASSOCIAZIONE: è una relazione conoscibile fra 2 o più variabili (fattori che variano fra le popolazioni e le retroazioni). Una associazione positiva si verifica quando l'aumento di una retroazione aumenta l'aumento dell'altra variabile (fumo - cancro polmonare). È negativa se all'aumento dell'una diminuisce l'altra.

CAUSALITÀ

George controfattuale: condizionale; corrisponde al periodo ipotetico
della realtà

Sia $Y_{X=1}$ la verosimile risultato che sarebbe stata osservata
sotto il valore di esposizione $X=1$ (trattato) e $Y_{X=0}$ la verosimile
risultato che sarebbe stata osservata sotto il valore di esposizione
 $X=0$ (non trattato).

Le variabili $Y_{X=1}$ e $Y_{X=0}$ sono note come potenziali risultati
perché seguono di loro deriva il valore dell'esito del soggetto
che sarebbe stato osservato sotto un potenziale scenario
de cui però egli non ha effettiva esperienza. Sono conosciuti come
ESITI CONTROFATTUALI. Per ogni soggetto, uno degli esiti controfattuali
è in realtà quello che corrisponde al regime di trattamento che
lui ha effettivamente ricevuto. Con la modellazione vediamo se

$$\Pr(Y_{X=1}=1) = \Pr(Y_{X=0}=1)$$

Sviluppo dei metodi statistici appropriati (MSK, DAG, IV, -)

Nel campo delle scienze sociali, un approccio empirico alle relazioni causali consiste nel guardare $\Delta X_t \rightarrow \Delta Y_{t'}$ a condizione che
un cambiamento nella variabile X al tempo t sia causa di un
cambiamento nella variabile Y in un secondo tempo t' (esso non
esiste che X sia l'unica causa di Y).

Requisiti:

- 1) momento dell'ordine fra causa ed effetto, $t < t'$
- 2) causalità può essere tempo dipendente, e le relazioni
può variare nel tempo
- 3) i singoli dati longitudinali, ossia i dati della cronologia
degli eventi, forniscono informazioni appropriate.

Per via delle casualità dovute a scelte individuali, ad errori
di misura o relazioni causali complesse, e fattori non misurati,
 $\Delta X_t \rightarrow \Delta P_t(\Delta Y_{t'})$ un cambiamento nella variabile X al tempo t
nella t' la probabilità che la var. dipendente Y cambi in futuro
($t < t'$).

L'effetto causale è la propensione a cambiare comportamento
sociale per ogni individuo.

ANALISI DELLA STORIA DEGLI EVENTI

es: analisi delle storie coniugali

Abbiamo bisogno di definire

1) Stato: le categorie della variabile dipendente (matrimonio,
sposato, divorziato, vedovo)

2) Evento: transizione da uno stato ad un altro da uno
stato di origine a uno stato di destinazione

(primo matrimonio, divorzio, non prima matrimonio, redovo)

3) durata: tempo trascorso fra un evento e un altro

4) periodo di rischio: periodo in cui qualcuno è a rischio di un particolare evento. Qualcuno può sperimentare il divorzio solo se sposato & un certo rischio in un'estate di tempo è formato da tutti i soggetti che sono a rischio di sperimentare l'evento in questione, in quell'estate di tempo.

* L'analisi delle stesse degli eventi è l'analisi della durata del non verificarsi di un evento durante il periodo di rischio

Se l'evento di interesse è il primo matrimonio, l'analisi riguarda la durata del non verificarsi del primo matrimonio cioè il tempo in cui gli individui sono rimasti nello stato di non essere sposati.

Eventi singolari (non ripetibili): prima nascita, primo matrimonio, morte, -

Eventi ripetibili: contratti di lavoro, n° di figlio, avesse, incidenti

• Modello a rischi competitivi: durante il corso della vita ogni soggetto è esposto simultaneamente alle azioni di molti altri rischi, relativi ad un evento (es decesso).

Tali rischi si presuffano competitivamente tra loro in quanto il manifestarsi di uno precluderà il verificarsi dell'altro. Quando la gente può passare da un stato di origine (nato) a uno di destinazione (nella analisi delle formazioni) nella prima unione può essere rilevante distinguere fra matrimonio e convivenza.

Per l'analisi dei tassi di mortalità si può desiderare di distinguere tra le cause di morte.

Modello a più stati: quando le persone possono muoversi altrove verso una sequenza di stati, gli eventi non possono essere caratterizzati solo dal loro stato di destinazione (come nel modello a rischi competitivo) ma differiscono anche in relazione al loro stato di origine (stato dell'occupazione: l'occupazione, la disoccupazione, fuori del lavoro, lavoro, -)

CONCETTI STATISTICI DI BASE

T : tempo del non verificarsi di un dato evento (ver reprobata)
(continuo o discratto)

$f(t)$: densità di T

$F(t)$: FdR di T

$S(t)$: funzione di sopravvivenza di T , probabilità dell'even.
verif. dell'evento prima di t

$h(t)$: funzione di rischio di T , rischio istantaneo di
verosimilità di un even. al tempo t , dato che l'even. non
si verifica prima di t .

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{d}{dt} F(t)$$

$$F(t) = \Pr(T \leq t) = \int_0^t f(u) du$$

$$S(t) = 1 - F(t) = \Pr(T > t) = \int_t^{+\infty} f(u) du = e^{- \int_t^{\infty} h(u) du}$$

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \left[\frac{\Pr(t \leq T \leq t + \Delta t | T > t)}{\Delta t} \right] = \left[\frac{\Pr(t \leq T \leq t + \Delta t \wedge T > t)}{\Pr(T > t) \Delta t} \right] = \\ &= \left[\frac{\Pr(t \leq T \leq t + \Delta t)}{\Pr(T > t) \Delta t} \right] = \frac{f(t)}{S(t)} = \frac{\frac{d}{dt} F(t)}{S(t)} = \frac{\frac{d}{dt}(1 - S(t))}{S(t)} = \\ &= - \frac{\frac{d}{dt} S(t)}{S(t)} = - \frac{d}{dt} (\ln S(t)). \end{aligned}$$

- $t > 0$

- può f, h, S , monotonie, discontinuus

- no probabilità, ma per t infinitesimo lo faccio

CENSURA

La censura si verifica quando il valore di una osservazione
è solo parzialmente conosciuto

t_1 : inizio dello studio

t_2 : fine dello studio

A: intervallo di censura

B: censura $\rightarrow x$

C: troncamento $\rightarrow x$

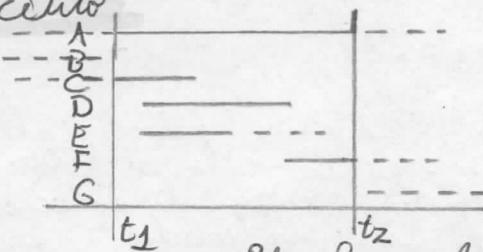
D: osservazione completa

E: censura amministrativa: x tra t_1 e t_2

F: censura completa

G: censura completa

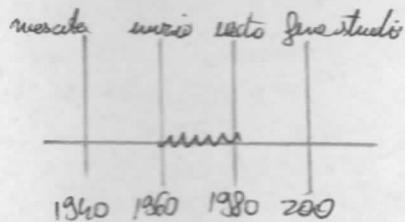
H: destra (solo inferiore)



I° tipo ferme lo studio se prima

II° tipo ferme lo studio dopo la obs.

SCALA TEMPORALE

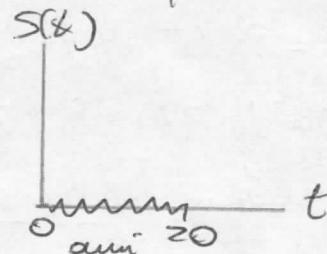


Possiamo scegliere due scale temporali

① scala del follow-up

è il tempo trascorso dalla scommessa nello studio. Per follow-up si intende un periodo di tempo successivo alla fine di un trattamento farmacologico o meno, durante il quale il soggetto è controllato periodicamente attraverso visite cliniche e esami strumentali.

È il follow-up di soggetti che hanno contrattato dipendentemente dalla età.

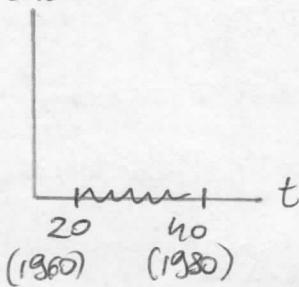


② scala d'età

è il tempo trascorso in anni di calendario

popolare a rischio: soggi che in un certo anno (1960) hanno una sex età (20 anni), dipende dall'età della popolazione

I soggetti a rischio sono quelli che hanno contrattato in quel intervallo di età.



Al variare della scala temporale variano i soggetti a rischio

GLHS (tempo)

Lo studio sulla storia delle vite tedesche (GLHS) comprende i dati sulle storie di vita di 8500 Mef provenienti da 20 Coorte. [Una coorte è un gruppo di persone identificate per un gruppo di caratteristiche. Uno studio di coorte è uno studio osservazionale che segue nel tempo l'evoluzione di una coorte. Può essere retrospettivo o prospettivo] di nascita in Germania-Ovest e da più di 2800 uomini e donne provenienti da 13 coorte di nascita in Germania Est. Gli intervistati sono stati tratti da campioni rappresentativi e sono stati analizzati in modo retrospettivo, riguardo la loro vita, in maniera tradizionale: faccio a faccio e interventi telefonici.

Particolare attenzione è stata dedicata al cambiamento nei modelli di extrusione e formazione, l'ingresso nel mercato del lavoro e dei processi di formazione delle famiglie e di come la trasformazione della società della Germania-Est avesse colpito i corsi di vita individuali.

Le domande vertevano sul tempo del verificarsi di determinati eventi, sulle condizioni storiche, il contesto estituzionale, sui comportamenti individuali, etc.

Per analizzare la durata del non verificarsi di un determinato evento è necessario conoscere:

- stato d'origine
- stato di destinazione
- tempo di avvio e di arresto

Attraverso i metodi parametrici e non parametrici si trova $h(t)$ ed $S(t)$

METODI NON PARAMETRICI

Questi metodi non fanno alcuna ipotesi circa la distribuzione del processo \rightarrow adatti per l'analisi esplorativa dei dati.

- tavola di mortalità
- metodo di Kaplan-Meier

Entrambi i metodi sono utili per il calcolo e la rappresentazione delle funzioni di sopravvivenza e del tasso di conversione $h(t)$

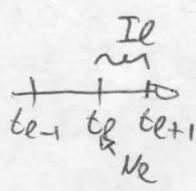
• Tavola di sopravvivenza: procedura tradizionale utilizzata nel caso di dati grandi poiché richiede meno tempo di calcolo e spazio, ma è necessario raggruppare le durate in intervalli fissi e i risultati dipendono quindi da quei intervalli di tempo definiti arbitrariamente. È necessario poter osservare in un relativamente elevato di eventi, in modo che le stime condotte per ogni intervallo siano affidabili.

• Metodo di Kaplan-Meier: non richiede di definire intervalli ma si basa sul calcolo di un rischio fissato in ogni istante di tempo in cui si è verificato almeno un evento

TAVOLE DI MORTALITÀ

$$0 \leq t_1 < t_2 < \dots < t_L \quad t_{L+1} = \infty \quad (t_0 = 0)$$

$$I_l = \{t \mid t_l \leq t < t_{l+1}\} \quad l = 1, \dots, L$$



E_l = n° di unità che subiscono l'evento in I_l

Z_l = n° di unità curvate in I_l

N_l = n° di unità che entrano in I_l

$R_l = \text{n}^{\circ} \text{ di unità le scivo a rischio di avere un evento in I}_l$

$$N_1 = N \quad N_l = N_{l-1} - E_{l-1} - Z_{l-1}$$

$$R_l = N_l - wZ_l \quad 0 < w \leq 1 \quad (w=1) \quad \text{protezione curvata che non contrabessa all'intervalle}$$

$$q_l = \frac{E_l}{R_l} \quad \text{probabilità condizionale di avere l'evento in I}_l$$

$p_l = 1 - q_l \quad \text{probabilità condizionale di sopravvivere in I}_l$

$$S_l = S_{l-1} \cdot p_{l-1} = \prod_{j=1}^{l-1} p_j = \prod_{j=1}^{l-1} \left(1 - \frac{E_j}{R_j} \right) = \prod_{j=1}^{l-1} \left(1 - \frac{E_j}{N_j - wZ_j} \right)$$

$$f_l = \frac{S_l - S_{l+1}}{t_{l+1} - t_l} \quad \text{calcolo nel punto medio dell'intervalle}$$

$$h_l = \frac{f_l}{\frac{S_l + S_{l+1}}{2}} \quad \text{funzione di rischio}$$

Se i campioni sono numerosi

$$\frac{\hat{S}_l - S_l}{S.B.(\hat{S}_l)} \sim N(0,1) \quad \frac{\hat{f}_l - f_l}{S.B.(\hat{f}_l)} \sim N(0,1) \quad \frac{\hat{h}_l - h_l}{S.B.(\hat{h}_l)} \sim N(0,1)$$

$$\Pr \left(Z_{\frac{\alpha}{2}} < \frac{\hat{S}_l - S_l}{S.B.(\hat{S}_l)} < Z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

Altrimenti

$$S.B.(S_l) = S_l \left(\sum_{j=1}^{l-1} \frac{q_j}{p_j R_j} \right)^{1/2}$$

$$S.B.(f_l) = \frac{q_l S_l}{t_{l+1} - t_l} \left(\frac{p_l}{q_l R_l} + \sum_{j=1}^{l-1} \frac{q_j}{p_j R_j} \right)^{1/2}$$

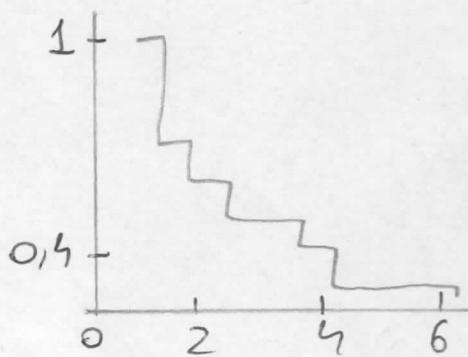
$$S.B.(h_l) = \frac{h_l}{q_l R_l} \left(1 - \left(\frac{q_l(t_{l+1} - t_l)}{2} \right)^2 \right)^{1/2}$$

ESEMPIO DI TAVOLA DI MORTALITÀ

Dato campione di 2418 uomini

te	I _l	N _l	E _l	Z _l	R _l	q _l	p _l	S _l
0	[0,1]	2418	- 456	0	2418	0,189	0,811	1
1	[1,2]	1962	226	39	1962,5	0,116	0,883	0,811
2	[2,3]	1697	132	22	1686	0,09	0,91	0,717
3	[3,4]	1523	171	23	1511,5	0,11	0,89	0,65
...								
9	[9,10]	427	42	64	385	0,106	0,894	0,33

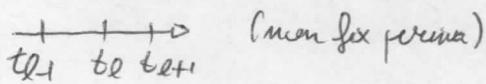
FUNZIONE DI SOPRAVIVENZA A TRAUBSO LA TAVOLA DI MORTALITÀ



Il problema di questo metodo è che è necessario definire a priori gli intervalli di tempo. Le stime sono affidabili solo se gli eventi osservati sono molti.

IL METODO DI KAPLAN-MEIER

Non richiede la definizione di $[t]$ perché calcola la $S(t)$ ogni volta che si verifica un evento.



N : unità entrate nello studio

E_l = numero di unità che subiscono l'evento in t_l

Z_l = numero di unità censurate in $[t_{l-1}, t_l]$

R_l = numero di unità che sono a rischio in t_l $= N_l$

Oss. $R_l = N_l$ (prima era ' $-w_l$ ') \rightarrow K-M NON CONSIDERA I CENSURATI COME 'OBEDENTI'. PRIMALI SONO TUTTI OAL OBNULENTE

$$\hat{S}(t) = \prod_{0 \leq t_l \leq t} \left(1 - \frac{E_l}{R_l}\right) = \prod_{0 \leq t_l \leq t} \left(1 - \frac{E_l}{N_l}\right)$$

Oss. Poiché è possibile che abbia I.C. $\circ e > 1$ (il tempo) allora trasferiamo la funzione di sopravvivenza: $\log(-\log(\hat{S}(t)))$

Così se $\hat{S}(t) = 0 \rightarrow$ dei +oo

se $\hat{S}(t) = 1 \rightarrow$ dei -oo

$$\begin{aligned} \text{Var}(\log(-\log(\hat{S}(t)))) &= \frac{1}{\log(\hat{S}(t))^2} \sum_{0 \leq t_l \leq t} \frac{E_l}{R_l(R_l - E_l)} \quad 95\% \text{ I.C.} \\ &= \log(-\log(\hat{S}(t))) - 1,96 \text{ s.e.} (\log(-\log \hat{S}(t))), \\ &\quad \log(-\log(\hat{S}(t))) + 1,96 \text{ s.e.} (\log(-\log \hat{S}(t))) \end{aligned}$$

Passaggi:

- Calcolo $\hat{S}(t)$
- Transform $\hat{S}(t)$
- calcolo I.C. (faccendo se non è simmetrico)
- poi ritrasform

STIMATORE DI NELSON - AALEN

Stimatore della funzione di rischio cumulativa $H(t)$

$$\hat{q}_e = \frac{\text{E}l}{R_e(t_{\text{fin}} - t_e)} \quad \text{n° eventi}$$

$$H(t) = \sum_{t \leq t_e < t} q_e (t_{\text{fin}} - t_e) \quad \left(\sigma_0 \int_0^t q(u) du \right)$$

$$\hat{H}_{NA}(t) = \sum_{t \leq t_e < t} \frac{\text{E}l}{R_e}$$

$$\hat{H}_{KH}(t) = -\log(\hat{S}_{KH}(t)) = -\log \left(\prod_{t \leq t_e < t} \left(1 - \frac{\text{E}l}{R_e}\right) \right) = -\sum_{t \leq t_e < t} \log \left(1 - \frac{\text{E}l}{R_e}\right)$$

$S(t) = e^{-\int_0^t q(u) du}$

Lo stimatore di N-A è anche stimatore della funzione di rischio cumulativa che si trova con K-H:

$$\begin{aligned} \hat{H}_{KH}(t) \rightarrow \hat{H}_{NA}(t) &= -\log(\hat{S}_{KH}(t)) = -\log \prod_{t \leq t_e < t} \left(1 - \frac{\text{E}l}{R_e}\right) = \\ &= -\sum_{t \leq t_e < t} \log \left(1 - \frac{\text{E}l}{R_e}\right) \approx -\sum_{t \leq t_e < t} \frac{\text{E}l}{R_e} = \hat{H}_{NA}(t) \end{aligned}$$

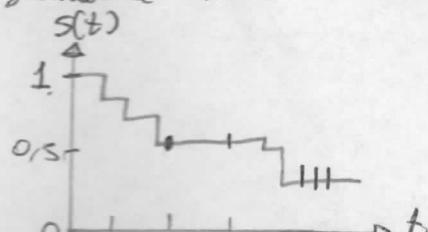
perché $-\log(1-h) \propto h$, la pendenza

Esempio

t_e	R_e	$\text{E}l$	Z_e	q_e	P_e	S_e	(formula $S_e = S_{e-1} - P_{e-1}$)
23	13	1	0	1/13	12/13	12/13	
47	12	1	0	1/12	11/12	12/13 · 11/12	
69	11	1	0	1/11	10/11	11/13 · 10/11	
70	10	0	1	0	10/10	10/13 · 1	
71	9	0	1				
100	8	0	1				
101	7	0	1				
148	6	1	0	1/6	5/6	10/13 · 5/6	
181	5	1	0	1/5	4/5	10/13 · 5/6 · 4/5	
198	4	0	1				
208	3	0	1				
212	2	0	1				
224	1	0	1				

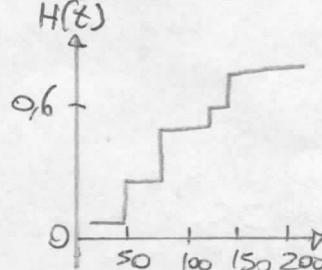
S_e non calcola le censure, se si calcola solo quel accade l'evento

Stimatore K-M



- la scalina = g(n° eventi)
- trattiui = censure

Stimatore N-A:



CONFRONTO FRA CURVE DI SOPRAVIVENZA

Per confrontare funzioni di sopravvivenza e testare se la differenza è significativa, possiamo calcolare intervalli di confidenza e valutare se si sovrappongono o meno. Oppure calcolare le statistiche dei test specifici per che due o più funzioni di sopravvivenza.

Sufficiamo di confrontare due gruppi independenti.

n° di eventi accaduti $E_{11}, R_{11}, E_{21}, R_{21}$ $E_{12}, R_{12}, E_{22}, R_{22}$

$E_{11} \quad n=3$
 $E_{12} \quad l=4$

Dove $t_{11}, t_{21}, \dots, t_{1l}$ sono istanti di tempo in cui è stato osservato almeno un evento nel campione

$$\begin{cases} H_0: S_1(t) = S_2(t) \\ H_1: S_1(t) \neq S_2(t) \end{cases}$$

$$Z = \frac{\widehat{S}_1(t) - \widehat{S}_2(t) - 0}{\text{se}(\widehat{S}_1(t)) + \text{se}(\widehat{S}_2(t))} \stackrel{H_0}{\sim} N(0, 1)$$

Questo test va fatto $\forall t \rightarrow$ test multipli

difficoltà \rightarrow proposto altro test

• Log-rank test: (o Mantel-Cox)

$$\begin{cases} H_0: S_1(t) = S_2(t) \\ H_1: S_1(t) \neq S_2(t) \end{cases}$$

se $0 < \theta < 1$, $S_1(t) > S_2(t)$
se $\theta > 1$, $S_1(t) < S_2(t)$
se $\theta = 1$, $S_1(t) = S_2(t)$

Siano E_{1e} e E_{2e} il numero di eventi osservato nei gruppi, rispettivamente al tempo l , con $E_l = E_{1e} + E_{2e}$. Pato che E_l eventi siano accaduti nei gruppi al tempo l , sotto H_0 , E_{1e} ha distribuzione ipergeometrica con valore atteso $A_l = E_l \frac{R_{1e}}{R_l}$ e varianza V_e con

$$V_e = \frac{E_{1e} R_{1e}}{R_l (1 - R_{1e}/R_l) (R_l - E_{1e})} \frac{R_l - 1}{R_l - 1}$$

$$Z = \frac{\sum_{l=1}^q (E_{1e} - A_l)}{\sqrt{\sum_{l=1}^q V_e}} \stackrel{H_0}{\sim} N(0, 1)$$

$$Q = \frac{\sum_{l=1}^q (E_{1e} - A_l)^2}{A_l} \stackrel{H_0}{\sim} K_1^2 \quad \sum_{l=1}^q \frac{(\text{val oss} - \text{val attesi})^2}{\text{val attesi}}$$

In generale, confrontando m grappi, $Q \sim K_m^2$

oss. Questo test è basato sul test- χ^2 di indipendenza: test non parametrico applicato quando i valori sono variabili nominali.

$t \in E_1, R_{11}$
 E_2, R_{12}

Tavella osservazioni
 soggetto Evento n° sognato ge 1

	E_1	$N_E - E_1$	
M 1	e_{11}	$n_{11} - e_{11}$	n_{12}
F 2	e_{12}	$n_{12} - e_{12}$	n_{22}
	e_2	$n_E - e_2$	n_E $n_{11} + n_{12}$

Tavella val. attesi ($+H_0$)

	E_1	$N_E - E_1$	
M 1	e_{11}	n_{11}	
F 2		n_E	

Sotto H_0 mi attendo di osservare questo: $\frac{e_{11} n_{12}}{n_E}$

Qui l'indipendenza è sufficente: ci sarebbe indipendenza fra la variabile sesso (M, F) e la variabile che indica la sopravvivenza

METODO PARAMETRICO

h(1)

I modelli sui taux de transition sono una base stabilita attraverso cui si può considerare come il tasso di transizione fra stati, e anche la funzione di sopravvivenza dipendono da un sistema di covariate (ma ci sono delle difficoltà nel quantificare le interpretazioni i risultati usando metodi non parametrici, soprattutto in presenza di molte variabili).

La versione più semplice è il contributo da oggi soggetto allo studio

Sono $i=1, \dots, n$ unità indipendenti

(t_i, J_i) è osservato, $t_i = \text{tempo del verificarsi dell'evento} \rightarrow \text{di curva}$
 $J_i = 1$ se è stato osservato l'evento
 $J_i = 0$ se si è osservata una censura

METODO MAX-LIKELIHOOD

$$L(\theta) = \prod_{i=1}^n \left[f(t_i; \theta) \cdot S(t_i; \theta)^{1-J_i} \right] = \prod_{i=1}^n h(t_i; \theta)^{J_i} S(t_i; \theta)^{1-J_i}$$

probabilità di un evento in $[t_{i-1}, t_i]$ condizionatamente ad esse sopravvissute fino a $t_i \rightarrow$ per ogni i risultato $J_i = 1$, è fatto svolgere f d' sopravvivenza complessa fino a t_i

$$\log L(\theta) = \sum_{i=1}^n \log [h(t_i; \theta)^{J_i} S(t_i; \theta)^{1-J_i}] = \sum_{i=1}^n \left[J_i \log h(t_i; \theta) + \log S(t_i; \theta) \right]$$

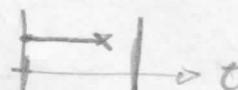
poi servo $\frac{\partial}{\partial \theta} \log L(\theta) = 0 \rightarrow$ ottengo θ .

Si cerca risposta a 'qual è l'effetto di una variabile sulla funzione di sopravvivenza?

moi de servir in ambito medico. (es. modelli con illo 4)
 solo x modellare se I dati

MODELLO ESPONENZIALE

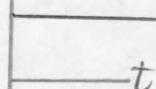
$T \sim e^\lambda \quad \lambda > 0$ tempo di sopravvivenza



$$f(t) = \lambda e^{-\lambda t}$$

$$S(t) = 1 - F(t) = 1 - \int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t} = e^{-\lambda t}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$



$$\begin{aligned} \log L(\lambda) &= \sum_{i=1}^m \left[\delta_i \log h(t_i, \lambda) + \log S(t_i, \lambda) \right] = \sum_{i=1}^m \delta_i \log(\lambda) - \lambda \sum t_i \\ &= \log(\lambda) \cdot \sum_i \delta_i - \lambda \sum_i t_i \end{aligned}$$

"event intensity"

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{\sum \sigma_i}{\lambda} - \frac{\sum t_i}{\lambda} = 0$$

Covariate

- età (\times studio morboso)
 - genere
 - livello di educazione

$$\hat{\lambda} = \frac{L}{8+1}$$

sogga \rightarrow contribuisce allo studio per 1 anno
soggb \rightarrow " " " " per 8 anni

le covariante sono costanti nel tempo:

$$h(t, x) = b(x) = \exp \left(\underline{\beta}^T x \right) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_K x_K}$$

$\underline{\beta}$ vector weights
+ intercept

- il tasso di recidiva ^{dei VTS, tra i soggetti} non dipende dal tempo
 - funzione di collegamento log-lineare tra il tasso di transizione e il vettore delle covariante
 - il tao di transizione è ≥ 0
 - questo modello permette di valutare l'effetto della variabile di interesse sul tasso di transizione, la presenza di fattori di confondimento, e interazioni fra le covariabili.

Spesso è difficile osservare un effetto delle correnti costante nell'arco temporale dello studio, vedo un altro modello.

MODELLO ESPONENZIALE COSTANTE A TRATTI^(no new selling)

In molte applicazioni, l'ipotesi che il tasso di crescita sia costante nel tempo non è giustificata e include quindi delle covariate tempo-dipendenti, suddividendo il tempo in periodi di tempo, assumendo che il tasso di crescita sia costante in ciascuno di questi intervalli, assumendo valori diversi (potenzialmente).

- * Solo la funzione di rischio baseale (e^{β_0}) può verificare tra i periodi di tempo, ma le covariate hanno gli stessi effetti in ogni periodo:

$$0 \leq t_1 < t_2 < \dots < t_L \quad t_{L+1} = \infty \quad (t_1 = 0)$$

$$I_L = \{t | t_L \leq t < t_{L+1}\} \quad L = 1, \dots, L$$

15

A. $h(t, x) = e^{\beta_0 t + \beta_1 x_1 + \beta_2 x_2 + \dots} \quad t \in I_l$

qui è introdotto una dipendenza nel parametrazione dell'intercetta, al vertice degli intervalli

$$\log L = \sum_{i=1}^n \left[\delta_i(t_i, l)(\beta_0 t + \beta^T x_i) - \sum_{l=1}^L \Delta(s_i, t_i, l) e^{\beta_0 t + \beta^T x_i} \right]$$

- $e^{\beta_0 t}$ non dipende dal tempo

• contributo i-esimo soggetto all'intervalle

• indica la sopravvivenza del soggetto ad un rischio

- $e^{\beta_1 x_1 + \beta_2 x_2 + \dots}$ è $\mu(x, \beta)$, predittore lineare

- una funzione di rischio baseline: $e^{\beta_0 t}$, come se fosse ' $\mu_0(t)$ ' [baseline]

$$\delta(t, l) = 1 \text{ se } t \in I_l, \text{ altrimenti } 0$$

s = tempo d'entrata

$$\Delta(s, t, l) = t - t_l$$

$$s \leq t_l, t_l < t < t_{l+1}$$

$$= t_{l+1} - t_l$$

$$s \leq t_l, t \geq t_{l+1}$$

$$= t_{l+1} - s$$

$$t \geq t_{l+1}, t_l < s < t_{l+1}$$

$$= 0 \text{ altrimenti}$$

B. • anche gli effetti delle covariate sono divisi nei vari periodi di tempo

$$h(t) = e^{\beta_0 t + \beta_1 l x_1 + \beta_2 l x_2 + \dots} \quad t \in I_l$$

$$\log L = \sum_{i=1}^n \left[\delta_i(t_i, l)(\beta_0 t + \beta^T l x_i) - \sum_{l=1}^L \Delta(s_i, t_i, l) e^{\beta_0 t + \beta^T l x_i} \right]$$

- introdotto dipendenza lineare di tutti i parametri da l

- buona performance del modello ma non è realistico

- problema di over-fitting \rightarrow troppi parametri da stimare \forall intervallo.

• periodi di tempo possono essere arbitrariamente definiti, ma:

- se si sceglie un gran numero di intervalli di tempo si ottiene una migliore approssimazione della funzione di rischio, riguardo, di base.

Ma ciò implica un gran numero di coefficienti da stimare

- se si sceglie un piccolo numero di periodi di tempo, ci sono meno problemi di stima, ma ci è una approssimazione più povera della funzione di rischio di base

- un compromesso è necessario: dovrebbero essere alcuni dati all'interno di ciascun intervallo di tempo.

Ma il modello esponenziale è poco realistico, per via del modello di Cox

GENERALIZZAZIONE

Gli intervalli di tempo non sono scelti arbitrariamente ma dipendono da un singolo individuo \rightarrow definizione di differenti intervalli per ogni soggetto:

Le covariate tempo-dipendenti potranno essere:

- definite come dipendenti dal tempo, il cui tempo totale percorso è determinato in antuario, allo stesso modo, per tutti i soggetti dello studio (età) \rightarrow dipendono dai processi esterni.
 - covariate tempo dipendenti accettorie, le cui determinazioni dipendono da un processo stocastico esterno all'unità oggetto di studio (tasso di contaminazione) e dipendono da processi stocastici.
 - covariate tempo dipendenti interne, le cui determinazioni dipendono dal soggetto dello studio
- Se le covariate tempo dipendenti sono quantitative, esse conterranno il loro valore in istante di tempo dato. In ogni istante di tempo, quando almeno una covariata cambia il suo valore, la situazione originale viene suddivisa in parti
 \rightarrow approssimazione dei cambiamenti della variabile quantitativa.

(considerando la situazione originale, fra le molte (forse il tempo) possibili, (j, N, s, t) con origine iniziale fine t , e stati intermedi j e k . Questa situazione si può dividere in L sottosituazioni (j_l, N_l, s_l, t_l) , $l=1, \dots, L$

$$h(t) = e^{\beta_0 t + \beta_1 l x_1 + \beta_2 l x_2 + \dots} \quad t \in I_l$$

$$S(t) = \prod_{l=1}^L S(t_l | s_e)$$

$$S(t_l | s_e) = e^{-\int_{s_e}^{t_l} h(u) du}$$

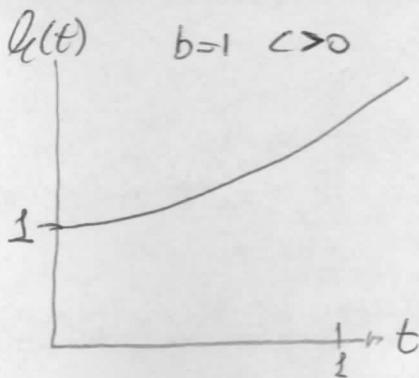
$$\begin{aligned} \log L &= \sum_i^n \left[J_i(t_0, l) \log h(t_0) + \sum_{l=1}^L \log S(t_l | s_e) \right] = \\ &= \sum_i^n \left[J_i(t_i, l) \log h(t_i) - \sum_{l=1}^L \int_{s_e}^{t_l} h(u) du \right] = \\ &= \sum_i^n \left[J_i(t_i, l) \log h(t_i) - \sum_{l=1}^L \int_{s_e}^{t_l} e^{\beta_0 u + \beta_1^T x_i u} du \right] \end{aligned}$$

MODELLO DI GOMPERTZ

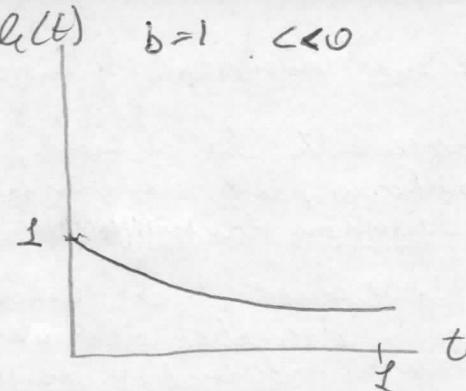
$$h(t) = b e^{ct} \quad b \geq 0$$

$$S(t) = e^{-\int_{s_e}^t h(u) du} = e^{-\int_{s_e}^t b e^{cu} du} = e^{-\frac{b}{c}(e^{ct} - 1)}$$

$$f(t) = S(t) h(t) = b e^{ct} \cdot c e^{(\frac{b}{c} e^{ct} - 1)}$$



- monotona crescente



- monotona decrescente

Se $c=0 \rightarrow$ modello esponenziale

Nei svolgimenti il tasso di uscita dal lavoro è guardiamo la durata (del lavoro) come aumento dell'abilità che si acquisisce in ogni nuovo lavoro: l'esperienza del lavoro comincia con ogni nuovo lavoro e cresce linearmente con il tempo trascorso in un posto di lavoro. (esperienza lavorativa cresce linearmente nel tempo)

- Modello senza covariate: $b = e^{\beta_0}$, $c = \gamma_0$
Supponiamo ipotizzare che con il aumento dei posti di lavoro specifici per la manodopera, la funzione di residenza decresce monotonicamente ($c < 0$)

Poiché se suppone che gli individui sonoeterogenei, si utilizzano le covariate (per modellare heterogeneità fra individui)

- Modello con le covariate legate a b

$$b = e^{\beta_0 + \beta^T X}, \quad c = \gamma_0$$

- Modello con le covariate legate ad entrambi: (b, c)

$$b = e^{\beta_0 + \beta^T X_1}, \quad c = \gamma_0 + \gamma^T X_2$$

MODELLO DI WEIBULL

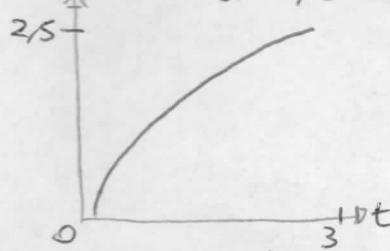
- molto usato perché molto flessibile

$$h(t) = b a^b t^{b-1} \quad a, b > 0$$

$$S(t) = e^{-\int_0^t h(u) du}$$

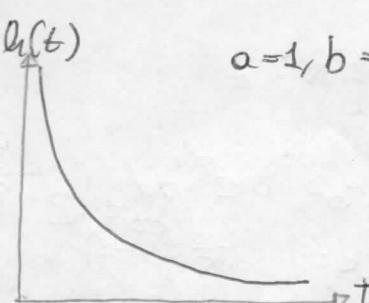
$$f(t) = b a^b t^{b-1} e^{-(at)^b}$$

$$h(t) \quad a=1, b=1,5$$



- crescente se $b > 1$

$$h(t) \quad a=1, b=0,5$$

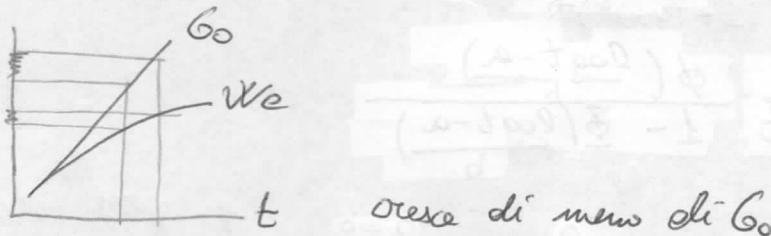


- decrescente se $0 < b < 1$

Quando una funzione di residuo è crescente o decrescente nel tempo si utilizza questo modello.

• È un modello che è detto a residuo proporzionali ed anche tempo accelerati, per la sua flessibilità.

• Frequentemente l'esperienza lavorativa qui è assente del tempo linearmente, ma logaritmicamente rispetto alle duree degli lavori.



Passo introdotto un modello con:

- nessuna covariata:

$$a = e^{-\alpha_0}$$

$$b = e^{\beta_0}$$

- con le covariate: modello con covariate legate al parametro a

$$a = e^{-\alpha_0 - \alpha^T x}$$

$$b = e^{\beta_0}$$

- con le covariate legate sia ad a che ad b:

$$a = e^{-\alpha_0 - \alpha^T x_1}$$

$$b = e^{\beta_0 + \beta^T x_2}$$

MODELLO DI GOMPERT E DI WEIBULL

Questa chiamata $x_j(t)$ una variabile latente che aumenta al passare del tempo t. Passo avere che:

$$\text{Co. } h(t) = e^{x\beta + x_j(t)\beta_j}$$

f di residuo cresce linearmente risp x_j

$$\text{We. } h(t) = e^{x\beta + \log(x_j(t)\beta_j)}$$

" " " " logaritmo risp x_j

x_{molti}

Lo qst variabile incide in due diversi modi sulla funzione di residuo

↳ Seguiamo una fra le due situazioni, a scende di come presta la crescita & abbilità del lavoro

La scelta fra Gomp e Weibull dipenderà dell'analisi dei residui, ma anche in maniera logica (ipotesi una nell'abbilità)

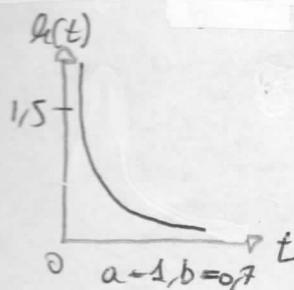
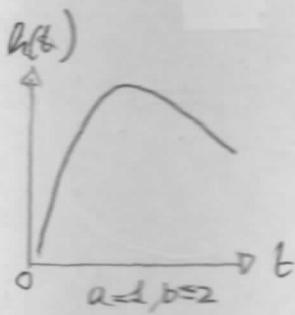
MODELLO LOG LOGISTICO

no farsi retti a memoria

$$h(t) = \frac{ba^b t^{b-1}}{(1+(at)^b)} \quad a, b > 0$$

$$S(t) = \frac{1}{1+(at)^b}$$

$$f(t) = \frac{ba^b t^{b-1}}{(1+(at)^b)^2}$$



$\Rightarrow b > 1 \rightarrow$ crescente e compattata
 $\Rightarrow b < 1 \rightarrow$ decrescente

È un modello ancora + flessibile: decrescente o compattato;
 Ciò permette di modellare fenomeni che Gomp. e Weib non
 potrebbero modellare, che hanno un basso livello crescente e poi un
 decrescente nel tempo. Es. basso di abbondanza nel lavoro: all'inizio
 è alto perché si cerca il lavoro giusto, una volta trovato il
 posto scende scende, quindi all'aumentare dell'esperienza lavorativa

Modello senza covariate:

$$\begin{aligned} a &= e^{-\alpha_0} \\ b &= e^{-\beta_0} \end{aligned}$$

esprimere a e b come \exp perché devono essere > 0

Modello con covariante legata ad a :

$$\begin{aligned} a &= e^{-\alpha_0 - \underline{\alpha}^T \underline{x}} \\ b &= e^{-\beta_0} \end{aligned}$$

Modello con covariante legata ad a e $a'b'$

$$a = e^{-\alpha_0 - \underline{\alpha}^T \underline{x}_1}$$

$$b = e^{-\beta_0 - \underline{\beta}^T \underline{x}_2}$$

MODELLO LOG-NORMALE

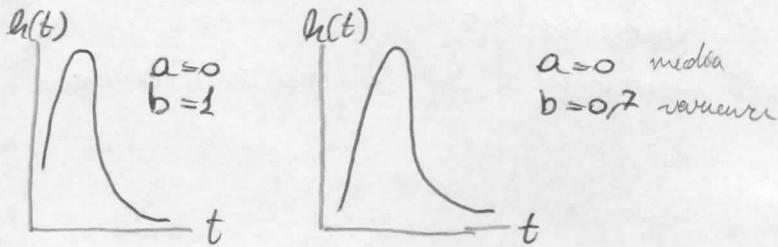
$\log T \sim N(\alpha, b)$ assumendo il log del tempo \sim Normale

$$h(t) = \frac{1}{bt} \frac{\phi(\frac{\log t - \alpha}{b})}{1 - \phi(\frac{\log t - \alpha}{b})} \quad b > 0$$

$$f(t) = \frac{1}{bt} \phi\left(\frac{\log t - \alpha}{b}\right)$$

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad \phi(t) = \int_0^t \phi(u) du$$

$$S(t) = 1 - \frac{1}{2} \left(\frac{\log t - a}{b} \right)$$



Questo modello può assumere solo una forma a campane.

Modello senza covariate:

$$a = \alpha_0$$

$$b = e^{\beta_0}$$

modello con covariante (per modellare l'eterogeneità fra gli individui) legate ad a

$$a = \alpha_0 + \underline{\alpha}^T \underline{x}$$

$$b = e^{\beta_0}$$

L modello log-logistico e log-normale sono i più usati per descrivere fenomeni la cui sfida risulta l'asymmetria (matremosità).

DISCUSSIONE SULLE COVARIATE

Prima di vedere i metodi per il controllo delle ipotesi ipocentesi, osserviamo alcune cose sulle covariante:

Tessato un modello che si adatta bene ai dati attraverso vari metodi e è introdotta una covariante, inserita in un parametrazione o in entrambi, risolve il problema di interpretazione dei系数 dei beta: β

• Una covariante può essere un fattore di confondimento:

es. alcool \rightarrow cancro ai polmoni
 fumo \rightarrow cancro ai polmoni
 assoc. fumo \rightarrow assoc. fumo

(NB) L'effetto dell'alcool sul cancro è costante al variare del fumatore

L'associazione fra alcool e cancro è forte, è confermata dal fumo

Se deve controllare per un fattore di confondimento:

① inserendolo in un modello di regressione come covariante

$$h(t, x) = h(t) e^{(x\beta_1)}, \quad x = \text{alcool} \quad \text{e } \beta_1 \text{ sarà significativo}$$

$$h(t, x) = h(t) e^{x_1 \beta_1 + c \beta_2}, \quad c = \text{confondente}$$

$h(t, x)$ è un modello gls, Normale, Cox, ...

2) Stratificando per il fattore di confondimento:

Si vede l'effetto dell'alcol sull'insorgenza del cuore sui problemi

3) Facendo studi di matching:

L'appaiamento toglie il confondimento. Sono cioè fra due gruppi di pazienti che sono appaiati per determinate caratteristiche; si distinguiscono in maniera simile per le covariante di sforzamento.

Qs) Gli studi che hanno meno confondimento di tutti sono i clinici, tra cui sono studi randomizzati. Si prendono a caso 2 gruppi, i trattati e non, che hanno $f(x)$ simili.

STUDIO CLINICO CONTROLLATO E RANDOMIZZATO: In uno studio clinico controllato i partecipanti sono assegnati in modo casuale a ricevere il trattamento sperimentale o il trattamento di controllo. Quando gli studi randomizzati sono condotti in modo appropriato, l'effetto del trattamento può essere studiato in gruppi di persone che sono:

- simili all'inizio
- trattate allo stesso modo, eccetto per intervento in studio.

Quindi qualsiasi differenza vera alle fine nei gruppi può essere attribuita esclusivamente al trattamento e non ad errori sistematici o al caso.

Qs) non potendo intervenire nel desegno dello studio perdeci e già pronto, con i dati, controllare il confondimento in fase di analisi

• Una covariata può essere un modificatore d'effetto:

a) genere \rightarrow divorzio
età

L'età modifica l'effetto del genere sul rischio di divorzio: somma il contributo dell'età; è l'interazione

L'età non è solo associata; gli uomini fra 35-40 anni hanno un elevato rischio di divorziare rispetto alle donne

Mentre prima seppurro che l'effetto dell'alcol sul formazione dei tumori fosse indipendente dal ~~fatto di essere femminile o non femminile~~ ^{CONFONDENTE}, qui invece l'effetto del genere ~~conferma~~ ^{MODIFICATORE D'EFFETTO} al contribuire dell'età, MA NON è costante.

L'interazione dell'età la scriviamo così: $c^{\beta_1 \text{genere} + \beta_2 \text{età} + \beta_3 \text{genere} \cdot \text{età}}$

COME SAPERE SE UN COEFF. È SIGNIFICATIVO O NO

ci sono diversi metodi, ne vediamo.

- Andando a confrontare questo modello, quello breve e quello col possibile fatto confondente.

mento:

$$\textcircled{1} h(t) = e^{\beta_0 + \beta_1 x_1} \quad (\text{base})$$

$$\textcircled{2} h(t) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

Sai confondere

Si vengono a vedere cose: $\hat{\beta}_1$; dice l'effetto della variabile sullo $f(x)$. (importante $\hat{\beta} \neq 0$)

• IC $\hat{\beta}$ al 95% (no p-value)

(NB) Se IC contiene lo zero, il test non ha un effetto statisticamente significativo (equivale a p-value > 5%)

Quando introduco il possibile fattore di confondimento vedo se il coefficiente dell'altra variabile (β_2) cambia: se cambia la stima di β_1 , allora è evidente che x_2 è fattore di confondimento. Dunque β_2 potrebbe anche essere non significativo, ma deve restare nel modello.

Arrivati qui:

• non usare la tecnica step-wise: prende se base del p-value vedere tutto esso in base a una regola che gli alle fine dà un modello contenente le variabili significative al 95%

→ So vedo se il $\hat{\beta}$ è significativo

→ Da stata vedo il test di Wald: $\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$

$$\frac{\hat{\beta} - 0}{SE(\hat{\beta})} \stackrel{H_0}{\sim} N(0,1)$$

• refuso se il valore sta sulle code

CRITERI PER CONFRONTARE MODELLI

• LRT - test raffigurazione

$$\frac{(\hat{\beta} - 0)^2}{SE(\hat{\beta})^2} \stackrel{H_0}{\sim} \chi^2_{k_1}$$

$$2[\log L_2 - \log L_1] \sim \chi^2_{k_2 - k_1} \quad \text{non parametrici}$$

• si vede se è significativo → se non lo è si usa un modello più semplice

• AIC: (akaike information criterio) (+tasso è meglio)
vale solo per modelli annidati uno con l'altro

$$2N - 2 \log L$$

\uparrow n° parametri
 \uparrow versione

• è un metodo che penalizza l'uso eccessivo di parametri nel modello quindi a parità di L preferisce il modello con meno parametri

$$\hookrightarrow \text{AIC} + \text{fattore}$$

• BIC

per modelli non annidati facoltà

Una volta stimato il modello, cosa è l'HR, mi interessa vedere

* L'effetto della covariata sulla funzione di rischio.

Hazard ratio: rapporto fra i rischi

$$HR = \frac{h(t, x^*)}{h(t, x)} = \frac{h_0(t)}{h_0(t)} e^{\sum \beta_i x_i^*} = e^{\sum \beta_i (x_i^* - x_i)}$$

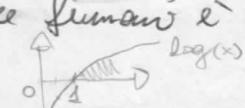
o rapporto fra i rischi di insorgenza di cancro nei fumatori e non fumatori (fra chi tiene e non)

- $HR = \frac{h(t, x=1)}{h(t, x=0)}$ X dichiarazione $x = \begin{cases} 1 & \text{fumatore} \\ 0 & \text{non fumatore} \end{cases}$
 nel modello
 $= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \Rightarrow \boxed{\beta_1 = \log HR}$

NON DI PENSARE
DALTBHPO

$$y = \beta_0 + \beta_1 x$$

logaritmo del
rapporto fra i
rischi
(NUMBER)



Se $\beta_1 > 0$, allora so che il rischio per coloro che fumano è > del rischio per coloro che non fumano $\rightarrow HR > 1$ (positivo)

Se $\beta_1 < 0$, allora il rischio per i fumatori è minore di quello per i non fumatori $\rightarrow HR < 1$ (negativo)

Se il coeff ha un effetto negativo $\rightarrow HR \in [0, 1]$
 " " " " " " positivo $\rightarrow HR > 1$

Nelle terine riguardano i coeff come HR:

- $\boxed{x \geq 1 : \text{essere fematori} \wedge \text{rischio}}$
- $\boxed{x=0 : \text{essere fematori} \wedge \text{rischio}}$

• HR

X discrete (livello educazione: 0 medie, 1 superiori, 2 laurea)
 qui devo scegliere la categoria di riferimento, es $x=0$

$$\left\{ \begin{array}{l} 2 \quad \frac{h(t, x=2)}{h(t, x=0)} = \exp(\beta_2) \\ 1 \quad \frac{h(t, x=1)}{h(t, x=0)} = \exp(\beta_1) \\ 0 \end{array} \right.$$

solo 2 rapporti di rischi; ormai il rischio per entrambi i gruppi rispetto alla categoria $x=0$

• HR

X continua

$$h(t) = \exp(\beta_0 + \beta_1 x) \rightarrow \boxed{\text{I categorie baseline quando considero l'effetto dell'incremento unitario di } x \text{ (}} = e^{\beta_1} \text{)}$$

qui β_1 è l'effetto dell'incremento unitario di x su $h(t)$

$$\frac{h(t, x+1)}{h(t, x)} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1} / \boxed{\text{effetto incremento unitario di } x \text{ sul rapporto}}$$

Modello esponenziale

⊕ senza covariante

$$h(t) = \lambda$$

$$h(t) = e^{\beta_0}, \beta_0 = \text{costante del modello}$$

t	coeff	s.e.	z	...
cons	-4,489	0,46	-9,6	...

ottenuta con Rho fit

$$\text{il tasso di rest. costante nel tempo, è } e^{-\beta_0} = 0,112$$

⊕ covariate costanti:

$$h(t) = e^{\beta_0 + \beta_1 \cdot x} = \lambda(t)$$

ognuno di qst va elevato alla exp

t	coeff	s.e.	z	...
edu	0,015	0,211	0,71	...
cons	-4,65	0,24	-19,33	

0,015 è l'effetto della variabile 'x' sul log h(t)

perche' ha + segno perciò gli HR per cui c'è effetto di un aumento numerico nelle var. cont.

$$\frac{\lambda_2}{\lambda_1} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1} = 1,1015$$

con il rapporto tra i risultati è di poco più di 1 quindi la variabile x aumenta di una unità.

(usa comando rho fit)

t	Hashtatton	s.e.
edu	1,015	0,211

MODIFICAZIONE DI EFFETTO

$$h(t, x) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

interazione fra 2 variabili

$$x_1 = \begin{cases} 0 & \text{Fe} \\ 1 & \text{Me} \end{cases}$$

$$x_2 = \begin{cases} 0 & \text{eta} \leq 35 \\ 1 & \text{eta} > 35 \end{cases}$$

suppongo β_3 significativo

TH: di qui aumenta l'effetto dei maschi rispetto le femmine < 35 anni?

ho scelto uno dei 2 strati paralleli

$$\hookrightarrow \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

$$\frac{h(t, \text{maschi} \leq 35)}{h(t, \text{femmi} < 35)}$$

TH: di qui aumenta l'effetto dei maschi rispetto le femmine ≥ 35 anni?

$$\hookrightarrow \frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)}{\exp(\beta_0 + \beta_2)} = \exp(\beta_1 + \beta_3) = \exp(\beta_1) \cdot \exp(\beta_3)$$

In stata ho gli output in H-R quindi dovrò moltiplicare l'H-R dell'essere maschi sulle femmine con $\text{eta} \leq 35$ anni, per HR associato alle interazioni.

COVARIATE TEMPO - PREDICTION

Non sempre le covariate sono costanti nel tempo

- Sono definite se variano allo stesso modo per tutti i soggetti (es. eta)
 - Sono casuali se variano nel tempo a causa di processi che sono indipendente dal singolo soggetto, ma dipendente da quegli esterni. (es. taz desoccupazione)
 - Sono interne se sono dipendenti dal soggetto (AIDS, assunz. farmaci)
- Le covariante tempo-dipendenti possono essere quantitative e qualitativa.

MODELLO A TEMPI ACCELERATI

Riguarda che puoi scrivere il tempo T in funzione del tempo baseline

Tutti i modelli introdotti finora sono modelli a tempo accelerato: si modella T condizionatamente a X del punto esprimere come un tempo T_0 se un predittore lineare.

$$T|X = \frac{T_0}{M(\underline{x}, \beta)} \quad \text{con} \quad T_0 = T|\underline{x}_0$$

\underline{x}_0 : vettore delle covariate base
 $M(\underline{x}, \beta) > 0$ predittore lineare

- se $\mu(\underline{x}, \beta) > 1$ le variabili esplicative riducono i tempi di sopravvivenza
- se $\mu(\underline{x}, \beta) < 1$ le variabili allungano i tempi di sopravvivenza

↳ dunque modello $T|X$ riducebbe o aumenterebbe il tempo per un individuo baseline

Spesso $M(\underline{x}, \beta) = e^{\beta^T \underline{x}}$

L'effetto delle covariate è quello di cambiare la scala del tempo.

Come si comportano le funzioni (siti) rispetto ad un individuo baseline?

- $S(t|x) = \Pr(T > t | X) = \Pr\left(\frac{T_0}{M(\underline{x}, \beta)} > t | X\right) =$
 $= \Pr(T_0 > t \cdot \mu(\underline{x}, \beta) | X) = \frac{S_0(t \cdot \mu(\underline{x}, \beta))}{\text{funz. sopravvivenza individuo baseline}}$

la $f \& S$ di un soggetto con covariate \underline{x} è pari alla $f \& S$ di un individuo baseline nel tempo, non t , bensì $t \cdot \mu(\underline{x}, \beta)$ dunque un tempo ridotto o allungato. Quando la $f \& S$ è ridotta o maggiorata rispetto quella dell'individuo baseline

- $f(t|x) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{f_0(t) \cdot \mu(\underline{x}, \beta) \Pr(t \cdot \mu(\underline{x}, \beta) \leq T_0 \leq (t + \Delta t) \cdot \mu(\underline{x}, \beta))}{\mu(\underline{x}, \beta)} =$
 $= \mu(\underline{x}, \beta) f_0(t \cdot \mu(\underline{x}, \beta))$

- $h(t|x) = \frac{f(t|x)}{S(t|x)} = \frac{\mu(\underline{x}, \beta) f_0(t \cdot \mu(\underline{x}, \beta))}{S_0(t \cdot \mu(\underline{x}, \beta))} = \mu(\underline{x}, \beta) h_0(t \cdot \mu(\underline{x}, \beta))$
funzione rischio baseline

es. $h(t|x) \propto h_0(t \cdot \mu(\underline{x}, \beta))$

la funzione di rischio con covariate \underline{x} è pari al prodotto del predittore lineare e $f \& S$ rischio in un tempo $t \cdot \mu(\underline{x}, \beta)$

cfr con Cox.

es. Expon, Weib, log-log, log-norm.
di
 f a tempo accelerato.

METODI GRAFICI PER VERIFICARE IPOTESI PARAMETRICHE

È importante controllare empiricamente l'adeguatezza dei modelli su cui si basa l'inferenza ad esempio:

- **metodo grafici**: si confrontano trasformazioni di stime non parametriche di funzioni di sovravivente con le previsioni provenienti da modelli parametrici. Nello specifico, la γ -curva di questi afferelli misurano con una stima non parametrica di una funzione di sovravivente secondo $K-M$ (stimatore) e con la stima non parametrica della funzione cumulativa con lo stimatore di N-A
 - **residui e pseudo-residui**: essi sono calcolati entrofratti per valutare i problemi di distribuzione. I residui sono differenze dei valori osservati da una variabile dipendente dai valori stimati sotto l'assunzione di uno specifico modello. Quando la variabile dipendente non è osservata, come la funzione di rischio nei modelli senza discriminazione, sono estesi gli pseudo residui.

METODI GRAFICI

$$v-a \cdot T = S(t) \quad K-\mu \quad \xrightarrow{p} \text{mebede gema funzione}$$

\searrow
 $u(t) \quad N-A$
 $f(t)$

I) Modello esponentiale:

$s(t) = e^{-\lambda t}$, con λ funzione di rischio

$$[F(t) = 1 - e^{-\lambda t}]$$

$$\hookrightarrow \log s(t) = -xt$$

stimo $S(t)$ con $K=4$

Per questo $\ln \hat{S}(t)$ resta $t \rightarrow \infty$ sembra retta,
esempio questo modello
vale anche per II, III, IV).

II) Modello Merkury

$$S(t) = e^{-(at)}$$

$$\log S(t) = (at)^b \rightarrow \log(-\log S(t)) = b \log(t) + b \log(a)$$

per plotto $\log(-\log \frac{S(t)}{27})$ su $\log(t)$

III) Modello log-logistico

$$S(t) = \frac{1}{1 + (at)^b}$$

$$\hookrightarrow \log\left(\frac{1-S(t)}{S(t)}\right) = b \log(a) + b \log(t)$$

plotto $\log\left(\frac{1-S(t)}{S(t)}\right)$ vs $\log(t)$

IV) Modello log-normale

$$S(t) = 1 - \Phi\left(\frac{\log(t) - a}{b}\right)$$

$$\hookrightarrow \Phi^{-1}(1-S(t)) = \frac{\log(t) - a}{b}$$

plotto $\Phi^{-1}(1-S(t))$ vs $\log(t)$

così ha lo scopo di verificare l'adeguatezza del modello rispetto ai dati.

Scelto un modello con le opportune covariate, vedo l'analisi dei residui

RESIDUI STANDARDIZZATI

$$\text{da } T|x = \frac{T_0}{\mu(x, \beta)} \text{ ottengo } \log(T_i | X_i) = \beta^T x_i + \underbrace{\log T_0}_{\text{errore vero}} + \varepsilon_i$$

• come fosse una regressione

se $\log(T|x)$ dipende dalla distribuzione di $\log(T_0)$:

- se $\log T_0 \sim$ Weibull $\rightarrow T_0 \sim$ valori estremi
- se $\log T_0 \sim$ logistica $\rightarrow T_0 \sim$ log-logistico
- se $\log T_0 \sim$ normale $\rightarrow T_0 \sim$ log-normale

• la distribuzione di ε_i

A seconda del modello, anche gli errori avranno una certa distribuzione

Per ogni soggetto del campione ho:

$(t_{i,x}, T_{i,x}, x_i)$
 tempo ↑ covariate
 verifiche subscr. B(1) n° i
 evento n° 0
 censura

$$\text{So che } \log(t_{i,x} | x_i) = \beta^T x_i + \varepsilon_i$$

con $\varepsilon_i = \log(T_i)$

$$\hat{\varepsilon}_i = \frac{\log \frac{T_i}{\hat{\beta}_{\text{acc}} X_i}}{S_i}, \quad S_i = \text{parametro di scala}$$

RESIDUI STANDARD

Dalla distribuzione degli $\hat{\varepsilon}_i$ capisco la correttezza del modello scelto

es. Scelto modello log-log, con arte corretto \rightarrow se la distribuzione dei residui è logistica la mia distribuzione avrà i correttamente. Così lo verifico con K-M sui residui quando se la trasformata è una retta

RESIDUI DI COX SNELL

Si lasciano sulla funzione di residuo cumulato

C-S propongono di eliminare i c.d. "pseudo-residui" che sono

$$\hat{\varepsilon}_i = \underbrace{s_i \int_0^{T_i} \hat{\lambda}(u, x_i) du}_{H_{\hat{\lambda}}(t, x_i)} \quad i=1, \dots, N$$

$s_i = \text{tempo inizio dell'evento}$
 $t_i = \text{tempo di fine dell'evento}$

Se il modello è appropriato: $(\hat{\varepsilon}_i, \delta_i) \sim \exp(1)$ ed il grafico $-\log(\hat{S}_{\hat{\lambda}}(\hat{\varepsilon}_i))$ vs $\hat{\varepsilon}_i$, calcolato su un parametrizzazione dovrebbe essere approssimativamente lineare, passando per l'origine con pendente variabile.

Oss. Perché $\sim \exp(1)$?

$$\begin{aligned} & X \quad F_x(x) \\ & \circ S_x(x) = e^{-H_x(x)} \quad \rightarrow H_x(x) = -\ln S_x(x) = -\ln(1-F_x(x)) \\ & \circ \Pr(H_x(x) \leq y) = \Pr(-\ln(1-F_x(x)) \leq y) = \Pr(1-F_x(x) \geq e^{-y}) = \\ & \quad \text{FDR della f.d.r. cum} \\ & \quad = \Pr(F_x(x) \leq 1-e^{-y}) = 1-e^{-y} \end{aligned}$$

ricorre UNIFORM(0,1)

da qui, siccome gli ε_i sono una f.d.r. cumulata, so che

$\hat{\varepsilon}_i \sim \exp(1)$

RESIDUI MARTINGALA E OBVIANZA

(t_i, δ_i, x_i) si calcolano rispetto alle f.d.r. cumulate

$$\hat{H}(t_i, x_i) \quad i=1, \dots, N$$

• residui martingala: $\hat{Y}_{mi} = \hat{\delta}_i - \hat{H}(t_i, x_i)$ [osservato - atteso]

indicebole $\begin{cases} 0 & \text{OB} \\ 1 & \text{ES} \end{cases} \Rightarrow$ residuo cumulato in t_i

- se il modello è appropriato, il grafico di \hat{r}_{Mi} vs $e^{\beta^T x_i}$ dovrebbe essere approssimativamente lineare
- il grafico degli \hat{r}_{Mi} vs variabili non incluse nel modello possono suggerire potenziali relazioni fra funzione di rischio e queste variabili \rightarrow fa capo se vengono inserite dunque corrette nel modello

es. x_1, x_2 potenziali corrette i solo x_1 è corretta nel modello
res. martingala



se vedo che c'è una relazione fra i residui e la covariata non inserita nel modello, il grafico me lo suggerisce, (non una trasformazione) os qui $\rightarrow \log x_2$

- residui di deviazione: trasformazione sui residui martingala $| E(\hat{r}_{di}) = 0$

• se il modello è appropriato $\hat{r}_{di} \sim \text{White Noise } (0, 1)$

• verificare la presenza di outliers e sono incorretti "per loro" i residui non devono avere alcun trend nel plot D-plot per emergere pattern non metti nel modello

MODELLO A RISCHI PROPORTIONALI - semiparametrico (Cox)

$$h(t; \underline{x}) = h_0(t) \mu(\underline{x}, \underline{\beta})$$

Dunque generale modello a rischi proporzionali

es. Exponential, Gompertz, Weibull, Cox models

MODELLO DI COX

$$h(t; \underline{x}) = h_0(t) e^{(\underline{\beta}^T \underline{x})}$$

Rappresenta rischi relativi tra soggetti con diversi vettori di regressori è costanti.

$$\frac{h_0(t, \underline{x}^*)}{h_0(t, \underline{x})} = \frac{\mu(\underline{x}^*, \underline{\beta})}{\mu(\underline{x}, \underline{\beta})}$$

[poisma: $h(t, \underline{x}) = \mu(\underline{x}, \underline{\beta}) h_0(t) \cdot \mu(\underline{x}, \underline{\beta})$]

[ora: $h(t, \underline{x}) = \mu(\underline{x}, \underline{\beta}) h_0(t)$]

modello a tempo
accelerato
modello a rischi
proporzionali

specifica l'effetto delle covariate sulla f.d.r.

- (NB) La grossa t è che la funz. di rischio baseline del modello di Cox dipende solo dal tempo t e non dalle covariate \underline{x} (mentre precedente)

È un modello semiparametrico perché la f.d.r. baseline non viene specificata (nel modello a tempo accelerato la distribuzione degli errori $\log(T_t)$ può essere $N(0, 1)$, logistica, exp, ... \rightarrow calcolata in modo non parametrico)

Quindi "semiparametrico" perché:

$$h(t | \underline{x}) = h_0(t) e^{\underline{x}^T \underline{\beta}}$$

stimato non parametricamente

stimato in maniera parametrica; cioè si può sfruttare la parametricità sottostante

a causa di questo modo la funzione, non potrò stimare il β con la funzione di max verosimigli. come fatto per gli altri modelli; introduso allora la c.d verosimiglianza parziale.

- è chiamato modello semi-parametrico a causa della forma non specificata di $h_0(t)$

- le covariate hanno un effetto moltiplicativo su $h(t)$, esse possono indurre solamente uno spostamento proporzionale sulla funzione di rischio ma non possono cambiare la forma della curva

$$\frac{h(t; x_1)}{h(t; x_2)} = \exp((x_1 - x_2)\beta) \Rightarrow \ln \frac{h(t; x_1)}{h(t; x_2)} = (x_1 - x_2)\beta$$

Supponiamo $x=0,1$

$$\ln \frac{h(t; x_1=1)}{h(t; x_2=0)} = \beta \rightarrow \ln[h(t; x_1=1)] - \ln[h(t; x_2=0)] = \beta$$

VEROSIMIGLIANZA PARZIALE

è una modifica delle classiche verosimiglianze che è partendo $L(\beta) = \prod_i [f(t_i; \beta)^{\delta_i} S(t_i; \beta)^{1-\delta_i}] = \prod_i h(t_i; \beta)^{\delta_i} S(t_i; \theta)$

potrà $h = (t, x) = h_0(t) e^{x\beta}$ e $S(t, x) = e^{-\int_0^t h_0(u) e^{\int_u^t \beta} du}$, ho

$$L(\beta) = \prod_i [h_0(t_i) e^{(x_i\beta)}]^{\delta_i} e^{-\int_0^{t_i} h_0(u) e^{(x_i\beta)} du}$$

ma $h_0(t)$ non è specificato \rightarrow si trova il β attraverso le informazioni osservate dei dati, senza considerare $h_0(t)$

Scopo: Ottenere di verosimiglianza che dia stime consistenti e asintoticamente normali

$i=1, \dots, n$ soggetti
 t_1, \dots, t_K

$R(t_j)$ insieme dei soggetti a rischio in t_j

x_j covariante di un soggetto che subisce l'evento in t_j

$h_j(t) dt$: - $h_j(t)$ indice il rischio che soggetto subisce un evento in t_j , sapendo che finora non subisce un altro evento

- dt : dt mi permette di interpretare (congiuntivamente a $h_j(t)$) la f_d di rischio come probabilità

La probabilità che solo un individuo con valore di covariante x_j subisce un evento int t_j , dato l'insieme dei rischi che concernono tutti gli individui a cui potrebbe capitare

$$\frac{h(t_j; x_j) dt}{\sum_{l \in R(t_j)} h_l(t_l) dt} = \frac{\frac{h_0(t_j) e^{(x_j \beta)}}{\sum_{l \in R(t_j)} e^{(x_l \beta)}} dt}{\sum_{l \in R(t_j)} e^{(x_l \beta)} dt} = \frac{e^{x_j \beta}}{\sum_{l \in R(t_j)} e^{x_l \beta}}$$

Intuitivamente: probabilità che un soggetto int t_j diverso mediano delle probabilità degli individui con covariante x_l che subiscono l'evento in quello stesso tempo

$$L(\beta) = \prod_{j=1}^K \left(\frac{e^{x_j \beta}}{\sum_{l \in R(t_j)} e^{x_l \beta}} \right)$$

Così dimostro che se considero queste come verosimiglianze parziali, massimizzandole per β , ottengo gli stimatori per β , che cercavo.

- È detta "verosimiglianza parziale" perché non viene calcolata come al solito, ma è proporzionale alla probabilità condizionata di osservare un evento
- Le covariate tempo-dipendenti possono essere considerate aggiornando i valori (loro) ad ogni istante di tempo successivo.
- Questa verosimiglianza può essere trattata come una verosimiglianza standard: stima dei coefficienti, I.C., test Wald, LR test, -
- $\hat{\beta} = \max_{\beta} (L(\beta))$: le stime di β sono convergenti e assintoticamente normali
- Sorgono difficoltà se ci sono tempi di scadenza. Se ne sono state proposte molte approssimazioni: metodo di Breslow, di Efron, verosimiglianza marginale diretta, verosim. parziale esatta, etc.

FUNZIONE DI RISCHIO BASELINE

Assumendo che la funzione di rischio fra 2 eventi osservati è crescente:

$$h_0(t_j) = \frac{d_j}{(t_j - t_{j-1}) \sum_{l \in R_j} e^{x_l \beta}}$$

rapporto fra n. di eventi osservati in t_j diversi, come per la verosimiglianza, la Σ degli esponentiali, cioè fra i Σ degli eventi ancora a rischio int t_j per i ~~non~~ individui lasciati.

La f di r baselino si calcola attraverso lo stimatore di Breslow, in maniera non parametrica. Esso è uno stimatore equivalente a quello di N-A.

Ubbiato lo studio di Breslow mette per stimare la f d r
cumulata:

$$\hat{H}_0(t) = \sum_{j|t_j \leq t} h_0(t_j)(t_j - t_{j-1}) = \sum_{j|t_j \leq t} \frac{d_j}{\sum_{x \in R_j} e^{x \beta}}$$

$\sum_{x \in R_j} e^{x \beta}$
f d r
 t_j = due
a scadenza

ASSUNZIONE DI PROPORZIONALITÀ

È l'ipotesi supponibile, da cui deriva il nome del modello di Cox
'rischi proporzionali'.

E la caratteristica fondamentale di Cox: considero $\frac{h(t,x)}{h_0(t)} = \exp(x\beta)$

da cui $\log h(t,x) - \log h_0(t) = x\beta$

\neq \rightarrow l'ipotesi di proporzionalità
per la costante di proporzionalità
($e^{x\beta}$) che qui è $x\beta$ puramente
scalar è additiva

es. soggetto COVARI

$$x = \begin{cases} 0 \\ 1 \\ 2 \end{cases}$$

$$\frac{h(t,x=1)}{h(t,x=0)} = e^{\beta}$$

β rischia soggetto
è costante nel tempo

$$\frac{h(t,x=2)}{h(t,x=0)} = e^{2\beta}$$

• Graficamente, l'ipotesi di proporzionalità:

$$\log [H_i(t)] = \log [H_0(t)] + x_i \hat{\beta}$$

$$\log H_i(t) - \log H_0(t) = x_i \hat{\beta} - x_j \hat{\beta}$$

stimato con Breslow
rispetto alla cumulata
rischia se essere
decrescente non lo fa
per $H_i(t)$.

Si vede graficamente se le z curve sono costante: che
non si intersecano, e dicono $H_i(t)$

si vede così accetto
l'ipotesi



• Alternativamente verifica l'ipotesi di proporzionalità con K-M:

$$\text{tranne K-M } \log(-\log S_i(t)) - \log[-\log S_0(t)] = x_i \hat{\beta} - x_j \hat{\beta}$$

se le z curve sono $N \dots$ → accetto l'ipotesi di proporzionalità

RESIDUI

Sono residui fatti a posteriori

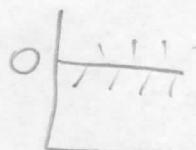
Il residuo di Schoenfeld sono definiti come la f.d.r. fra le covariate x_j di un soggetto che ha subito l'evento in t_j e la media delle covariate di un soggetto accaduto in t_j con peso proporzionale a $e^{(X\beta)}$:

$$r_j = x_j - \frac{\sum_{l \in R_j} x_l e^{x_l \beta}}{\sum_{l \in R_j} e^{x_l \beta}} = x_j - \hat{E}(x_0 | R_j)$$

$\sum_{l \in R_j}$ soggetti locoaccaduti

dice che è uguale alla media dei soggetti con covariabili x_l subite l'evento in t_j .

Se l'ipotesi di proporzionalità è corretta $\rightarrow E(r_j) = 0 \quad \forall t_j$



Così che se ripetiamo sul grafico nella stessa serie dei residui trasformati la $E(x_0 | R_j) = \text{stima del coeff } \beta$

Se $E(r_j) > 0 \rightarrow \text{HR aumenta}$

Se $E(r_j) < 0 \rightarrow \text{HR diminuisce}$

RESIDUI DI MARTINGALA:

$$\hat{r}_{m_i} = T_i - \hat{H}(t_i, x_i)$$

la f.d.r. stimata con Breslow
sono come quelli di fornia

RESIDUI SCORB:

confrontano i β stimati utilizzando tutti i dati con i β stimati senza utilizzare l'osservazione

- sono usati per valutare se c'è un outlier, o una osservazione che influisce molto sul trend del modello

- si vede da un grafico

- se si notano grosse differenze, si deve dimostrare quell'osservazione, a causa delle sue influenze.

MODELLO DI COX STRATIFICATO

Se l'ipotesi di proporzionalità non è valida (2 curve su 1) e res. Schoenfeld non vengono bene) altre volte si è dovuto ad una covariabile

sesto vce grafico allora si puo' fare un modello di Cox stratificato rispetto quella covariata

es. suppongo di volutare effetto trattamento

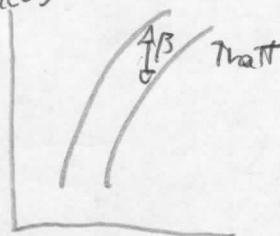
$$x_1 = \begin{cases} 1 & \text{Treat} \\ 0 & \text{NonTreat} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{Fem} \\ 0 & \text{Mas} \end{cases}$$

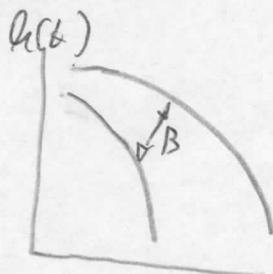
i 2 grafici di $\hat{S}(t)$, con K-M denonato.

suppongo che per il trattamento l'ipotesi di proporzionalita' e' OK, ma per il genere no

$$h_i(t) \quad \text{NonTreat}$$



Maschi: i & proporz.



Femmine: i & proporzionalità

TH: provare un modello di Cox stratificato per covariata-genere

Percio' si propone un $h_{ij}(t; x) = h_{0j}(t) \cdot e^{x_j B}$ (in alto costr.)

avremo una fdr & per M e F, purtroppo la differenza & solo nella fdr & baselina (stesso coeff di trattamento)

Guardando i grafici: il rapporto B e' sempre = sia nei M che F, una contro la fdr baselina (nisi M e' & nel tempo, nelle F & nel tempo)

Il modello di Cox stratificato dice che le fdr baseline differiscono per gli strati, ma la stessa doppia linea di rischio e lo stesso negli strati.

In sostanza dire se modello di Cox stratificato ha l'ip. soddisfatta o meno. Commentare.

• Usato molto in pratica, se l'ipot. e soddisfatta, per la semplicita' e l'interpretabilita'

All'interno di questo modello si possono introdurre delle variabili tempo-dipendenti. Sebbene il modello di G_x dipende dal tempo solo tramite la f_{dx} baseline, all'interno del predittore lineare posso introdurre variabili dipendenti dal tempo.

Ricordo che in G_x la f_{dx} baseline non dipende mai dalla covariata, al di fuori del termine tempo accelerato.

L'importante è non mettere nelle f_{dx} baseline le covariate.

PROGRAMMA DEL II° MODULO

Introduzione ai principali tipi di studio con le loro relative caratteristiche e diversità
Introduzione alla causalità

Definizione di "event history data" con relativi esempi

Definizione delle funzioni basilari la cui stima è richiesta nel corso di un'analisi, con relativi esempi

Metodi non parametrici utilizzati per stimare le funzioni d'interesse, quali la funzione di rischio, di sopravvivenza, di densità, etc, in particolare il metodo utilizzato per costruire una tavola di sopravvivenza, ed il metodo proposto da Kaplan-Meier

Confronto grafico tra molteplici curve di sopravvivenza e relativi test d'ipotesi

Covariate tempo-dipendenti con relativi esempi

Modello esponenziale, costante a tratti e no, modello di Gompertz-Makeham, modello di Weibull, modello log-logistico, modello log-normale, introdotti in molteplici contesti: senza covariate, con covariate costanti nel tempo, con covariate tempo-dipendenti, con un unico end-point d'interesse o con molteplici end-points d'interesse

Modello a rischi proporzionali con particolare attenzione alle ipotesi sottostanti l'assunzione del modello

Problemi inerenti alla scelta del modello ed alla sua interpretazione

Utilizzo del software Stata

TESTI CONSIGLIATI

Il modulo: Techniques of Event History Modeling: New Approaches to Causal Analysis.
2d ed. By Hans-Peter Blossfeld and Götz Rohwer. Mahwah, NJ: Lawrence Erlbaum, 2002.