

Metodologia di applicazione dei modelli di massima verosimiglianza per il trattamento dei dati missing

Erminio A. Bonizzoni

Congresso nazionale BIAS 29/30 aprile 2009, Sesto Fiorentino, Firenze

Stimatori di Massima Verosimiglianza

- ✓ Il principio alla base del metodo della **Massima Verosimiglianza** è la scelta di stimatori che **massimizzano la probabilità** di osservare ciò che abbiamo osservato nel nostro dataset
- ✓ È necessario conoscere a priori la **distribuzione di probabilità dei dati**: funzione sia dei dati stessi, sia dei parametri non-noti
- ✓ Condizionando per i dati osservati e massimizzando per i parametri non-noti si ottengono gli **stimatori di massima verosimiglianza**

$$\ell(\Theta; Y) = \prod f(y_i; \Theta) \leftrightarrow \ell\ell(\Theta; Y) = \sum \log f(y_i; \Theta)$$

$$\Theta_{ML} = \operatorname{argmax} \{ \ell\ell(\Theta; Y) \}$$

Massima Verosimiglianza e Dati Missing

- ✓ È il capofila dei metodi cosiddetti “principled” dal momento che costituisce il supporto teorico su cui poggia la classificazione dei dati mancanti in MCAR, MAR e MNAR
Tale classificazione è alla base dell’approccio sistematico alla gestione dei dati missing

ASSUNTO ESSENZIALE

Gli stimatori di massima verosimiglianza sono non-distorti quando il meccanismo che ha generato i dati mancanti è di tipo “Missing At Random” (MAR)

base di partenza (e sovente anche punto di arrivo)
di ogni analisi statistica in presenza di rilevazioni incomplete

Massima Verosimiglianza e MCAR

- ✓ Raramente i dati missing sono di tipo MCAR
- ✓ In genere esiste “almeno” una associazione fra la probabilità di missing e l’osservato
es. il trattamento ricevuto, i valori al basale e, negli studi longitudinali, le osservazioni che precedono l’evento di drop-out
- ✓ Un’analisi che include solo le osservazioni complete produrrà delle stime distorte

Massima Verosimiglianza e MAR

- ✓ In presenza di un meccanismo di tipo MAR, la probabilità che i dati siano missing spesso dipende dai valori (non-osservati) della variabile di risposta
- ✓ Tuttavia, condizionando per tutti gli altri valori (osservati), questa dipendenza viene interrotta e i dati missing diventeranno di tipo MCAR

$\text{Prob.}(y = \text{missing}) \rightarrow \text{MAR}$

$\text{Prob.}(y = \text{missing} \mid x, z, w, \dots) \rightarrow \text{MCAR}$

- ✓ Ciò implica che la distribuzione dei potenziali dati mancanti è la stessa per tutti i pazienti che condividono gli stessi dati osservati a prescindere dal fatto che presentino o non presentino dati missing

Massima Verosimiglianza e MAR

- ✓ In generale, negli studi longitudinali e in presenza di un meccanismo di tipo MAR, i soggetti drop-out condividono con i soggetti non drop-out lo stesso “trend” futuro (non osservato) dato il loro passato (osservato)
- ✓ Questa proprietà permette ai metodi *principled*, come quelli basati sulla verosimiglianza, di produrre delle stime non distorte in presenza di dati mancanti di tipo MAR



Missing At Random e Scenari Applicativi

- ✓ Il modo migliore per definire un modello statistico per dati mancanti consiste nel valutare la sua consistenza nell'ambito dei vari meccanismi di non-risposta (MCAR, MAR e MNAR)
- ✓ La base di partenza consiste sempre nell'assumere che il meccanismo che ha generato i dati mancanti è di tipo MAR
- ✓ Ciò significa che il modello statistico dovrà tener conto anche delle variabili predittive di drop-out
- ✓ Di conseguenza, è necessario, in prima istanza, identificare i predittori indipendenti di drop-out



Missing At Random e Scenari Applicativi

Negli esempi che seguiranno si assume che:

1. **i dati, originali o opportunamente trasformati, abbiano una distribuzione normale multivariata**
2. **il pattern dei dati mancanti sia di tipo monotono (drop-out) (per semplicità)**

1° Scenario Missing At Random (1)

- ✓ **Disegno Sperimentale:**
un solo trattamento e due visite al follow-up (y_{i1} , y_{i2}) in aggiunta al basale (y_{i0})

- ✓ **Meccanismo di Drop-out:**
la probabilità di drop-out può dipendere dal valore delle variabili di risposta rilevate ai tempi precedenti

$$\text{Prob.}(y_{i2}=\text{missing} \mid y_{i1}, y_{i0}) \rightarrow \text{Prob.}(y_{i1}=\text{missing} \mid y_{i0}) \rightarrow \text{MCAR}$$

- ✓ **Analisi di Massima Verosimiglianza:**
gli stimatori di massima verosimiglianza corretti per il Bias introdotto dal meccanismo Missing At Random, si ottengono dall'analisi simultanea dei seguenti tre modelli multivariati

1° Scenario Missing At Random (2)

$$(y_{i0}) \approx N\{(\mu_0), (\sigma_0)\}$$

$$\begin{pmatrix} y_{i0} \\ y_{i1} \end{pmatrix} \approx N\left\{ \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right\}$$

$$\begin{pmatrix} y_{i0} \\ y_{i1} \\ y_{i2} \end{pmatrix} \approx N\left\{ \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & & \\ \sigma_{01} & \sigma_1^2 & \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix} \right\}$$

$$(\mu, \Omega)_{ML} = \text{argmax}\{\ell\ell_1(\mu_0, \Omega_1) + \ell\ell_2(\mu_0, \mu_1, \Omega_2) + \ell\ell_3(\mu_0, \mu_1, \mu_2, \Omega_3)\}$$

1° Scenario Missing At Random (3)

- ✓ In pratica utilizzando la *Procedura MIXED* del SAS o altri software analoghi, non è necessario isolare e interpolare separatamente i modelli marginali per i tre diversi pattern di dati mancanti, ma è sufficiente fornire tutti i dati osservati in forma tabulare classica
- ✓ In sostanza il modello statistico non fa alcuna assunzione circa la funzione che ha generato i drop-out, tranne per il fatto che la probabilità di drop-out può dipendere dai dati osservati

2° Scenario Missing At Random (1)

- ✓ **Disegno Sperimentale:**
2 trattamenti e 1 visita al follow-up (y_{i1}) in aggiunta al basale (y_{i0})
- ✓ **Meccanismo di Drop-out:**
prevede che la probabilità di drop-out possa dipendere sia dal valore della variabile di risposta misurata al basale, sia dal trattamento
$$\text{Prob.}(y_{i1}=\text{missing} \mid y_{i0}, T) \rightarrow \text{MCAR}$$
- ✓ **Analisi di Massima Verosimiglianza:**
gli stimatori di massima verosimiglianza corretti per il Bias introdotto dal meccanismo Missing At Random, si possono ottenere in due modi

2° Scenario Missing At Random (2)

Primo metodo:

Analisi della Covarianza classica utilizzando i dati di tutti i pazienti con:

- a. risposta osservata al follow-up
- b. valori basali e trattamento come covariate

$$y_{i1} \approx N(\alpha_0 + \alpha_1 y_{i0} + \alpha_2 T, \sigma_{y_1}^2)$$

La stima della differenza fra trattamenti ottenuta in questo modo sarà:

- a. non distorta in caso di Missing At Random
- b. aggiustata per il basale (stima condizionata)

2° Scenario Missing At Random (3)

Secondo metodo:

Modello multivariato che include il basale come ulteriore variabile di risposta e il trattamento come covariata

Come nello scenario precedente, stimatori non distorti si possono ottenere dall'analisi simultanea dei seguenti due modelli statistici

$$y_{i0} \approx N(\alpha_0 + \alpha_1 T, \sigma_{y_0}^2)$$
$$\begin{pmatrix} y_{i0} \\ y_{i1} \end{pmatrix} \approx N \left\{ \begin{pmatrix} \alpha_0 + \alpha_1 T \\ \beta_0 + \beta_1 T \end{pmatrix}, \begin{pmatrix} \sigma_{y_0}^2 & \\ \sigma_{y_0 y_1} & \sigma_{y_1}^2 \end{pmatrix} \right\}$$

La stima della differenza fra trattamenti ottenuta in questo modo sarà:

- a. non distorta in caso di Missing At Random
- b. non aggiustata per il basale (stima marginale)

2° Scenario Missing At Random (4)

- ✓ Una variabile predittiva di drop-out deve essere inclusa nel modello, o come ulteriore variabile di risposta o come covariata
In entrambi i casi le stime ottenute saranno corrette per la componente MAR spiegata dalla variabile in questione
- ✓ Se si vuole stimare l'effetto del trattamento aggiustato (stima condizionata) per tale variabile, allora dovrà essere inserita nel modello come covariata
- ✓ Se si vuole stimare l'effetto del trattamento non aggiustato (stima marginale) per tale variabile, allora dovrà essere inserita nel modello come ulteriore variabile dipendente

3° Scenario Missing At Random (1)

- ✓ **Disegno Sperimentale:**
2 trattamenti, 1 visita al follow-up (y_{i1}) in aggiunta al basale (y_{i0}), 1 variabile basale predittiva di drop-out (x_{i0})
- ✓ **Meccanismo di Drop-out:**
prevede che la probabilità di drop-out possa dipendere sia dal valore della variabile di risposta misurata al basale, sia dal trattamento, sia da un'ulteriore variabile basale predittiva di drop-out (x_{i0})
$$\text{Prob.}(y_{i1}=\text{missing} \mid y_{i0}, T, x_{i0}) \rightarrow \text{MCAR}$$
- ✓ **Analisi di Massima Verosimiglianza:**
è necessario includere anche la variabile basale (x_{i0}) nel modello statistico per ottenere delle stime valide sotto ipotesi MAR

3° Scenario Missing At Random (2)

Analisi di Massima Verosimiglianza:

non si vuole condizionare la stima dell'effetto del trattamento per tali variabili, pertanto la variabile x_{i0} andrà ad aggiungersi alla variabile dipendente y_{i1} in un modello bivariato con il trattamento e il basale come covariate (oppure alle variabili dipendenti y_{i1} e y_{i0} in un modello trivariato con il solo trattamento come covariata)

$$\begin{pmatrix} x_{i0} \\ y_{i1} \end{pmatrix} \approx N \left\{ \begin{pmatrix} \alpha_0 + \alpha_1 y_{i0} + \alpha_2 T \\ \beta_0 + \beta_1 y_{i0} + \beta_2 T \end{pmatrix}, \begin{pmatrix} \sigma_{x_0}^2 & \sigma_{x_0 y_1} \\ \sigma_{x_0 y_1} & \sigma_{y_1}^2 \end{pmatrix} \right\}$$

3° Scenario Missing At Random (2)

Analisi di Massima Verosimiglianza

Nell' esempio precedente, la stima della differenza fra trattamenti (β_2) è:

- a. non distorta in caso di MAR
- b. aggiustata per il basale della variabile di risposta (y_{i0})
- c. non aggiustata per la variabile basale predittiva di drop-out (x_{i0})

Come negli scenari precedenti, la presenza contemporanea di risposte complete e parziali (per esempio, x_{01} non-missing e y_{i1} missing) non comporta alcun problema se si procede massimizzando direttamente la funzione di massima verosimiglianza utilizzando la Procedura MIXED del SAS o un software analogo

4° Scenario Missing At Random (1)

✓ **Disegno Sperimentale:**

2 trattamenti, 1 visita al follow-up (y_{i1}) in aggiunta al basale (y_{i0}), 2 variabili predittive di drop-out (x_{i0} -variabile Basale, z_{i0} -Variabile Post-Basale)

✓ **Meccanismo di Drop-out:**

prevede che la probabilità di drop-out possa dipendere dal basale, dal trattamento e da due ulteriori variabili predittive di drop-out, una misurata al basale e una rilevata alla visita di FU

$$\text{Prob.}(y_{i1}=\text{missing} \mid y_{i0}, T, x_{i0}, z_{i1}) \rightarrow \text{MCAR}$$

✓ **Analisi di Massima Verosimiglianza:**

per ottenere delle stime valide sotto ipotesi MAR, è necessario che tutti i predittori di drop-out siano parte del modello statistico o come covariate o come ulteriori variabili dipendenti (scenario precedente)

4° Scenario Missing At Random (2)

1. Basale della variabile di risposta:

è preferibile utilizzare il basale come covariata a prescindere dal tipo di meccanismo di drop-out, per poter ottenere una stima della differenza fra trattamenti aggiustata per il basale

2. Variabile basale predittore di drop-out:

generalmente non si ritiene di dover condizionare le stime della differenza fra trattamenti per queste variabili, quindi la loro collocazione più naturale è quella di ulteriori variabili dipendenti

3. Variabile post-basale predittore di drop-out:

l'unica collocazione possibile è quella di ulteriori variabili dipendenti (se si condiziona per variabili post-basali si rischia di mascherare gli effetti dei trattamenti sulla variabile di risposta di interesse)

4° Scenario Missing At Random (3)

Per questo scenario sono possibili tre modelli multivariati:

$$N(y_{i1}, z_{i1} \mid T, y_{i0}, x_{i0}, \Theta, \Omega)$$

$$N(x_{i0}, y_{i1}, z_{i1} \mid T, y_{i0}, \Theta, \Omega)$$

$$N(x_{i0}, y_{i0}, y_{i1}, z_{i1} \mid T, \Theta, \Omega)$$

Massimizzando le rispettive funzioni di massima verosimiglianza, si ottengono stime non distorte sotto ipotesi MAR

predittori di drop-out basali: è possibile condizionare e marginalizzare

predittori di drop-out post-basali: è possibile solo marginalizzare

5° Scenario Missing At Random (1)

Premessa:

gli esempi precedenti sono propedeutici allo scenario finale tipico di una sperimentazione clinica

✓ Disegno Sperimentale:

estensione dello scenario precedente al contesto longitudinale con più visite al follow-up

✓ Meccanismo di Drop-out:

sebbene l'analisi primaria è spesso un confronto fra trattamenti alla visita finale o a una visita prestabilita, quando i pazienti interrompono lo studio prematuramente, i dati sulla variabile di risposta raccolti prima dell'evento di drop-out conterranno verosimilmente un'elevata quantità di informazione sulle ragioni del drop-out stesso e per questo motivo sono considerati i più importanti predittori di drop-out

5° Scenario Missing At Random (2)

✓ Meccanismo di Drop-out :

per ottenere delle stime non distorte sotto ipotesi MAR, tutte le misure ripetute, e non solo l'osservazione finale, dovranno essere considerate nel modello statistico. Inoltre, dovranno necessariamente essere trattate come variabili di risposta in aggiunta all'osservazione finale

✓ Analisi di Massima verosimiglianza:

se si desidera condizionare per il basale della variabile di risposta e in presenza di altre variabili (basali e post-basali) predittori di drop-out, la funzione di densità di probabilità da cui trae origine la funzione di massima verosimiglianza sarà la seguente:

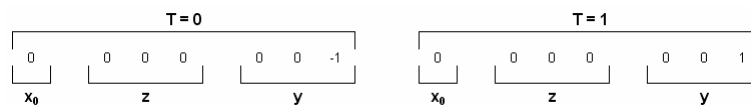
$$N(y_{i1} \dots y_{in}, z_{i1} \dots z_{in}, x_{i0} | T, y_{i0}, \Theta, \Omega)$$

5° Scenario Missing At Random (3)

ID	T	y ₀	new y	new T	Variabile
1	0	40,5	11,5	1	x ₀
1	0	40,5	0,8	2	z ₁
1	0	40,5	0,7	3	z ₂
1	0	40,5	1,1	4	z ₃
1	0	40,5	40,3	5	y ₁
1	0	40,5	59,1	6	y ₂
1	0	40,5	57,4	7	y ₃
2	1	39,0	12,0	8	x ₀
2	1	39,0	0,9	9	z ₁
2	1	39,0	0,9	10	z ₂
2	1	39,0	0,7	11	z ₃
2	1	39,0	45,0	12	y ₁
2	1	39,0	50,0	13	y ₂
2	1	39,0	60,0	14	y ₃
3	0	41,2	9,8	1	x ₀
3	0	41,2	0,4	2	z ₁
3	0	41,2	0,6	3	z ₂
3	0	41,2	0,4	4	z ₃
3	0	41,2	39,0	5	y ₁
3	0	41,2	47,0	6	y ₂
3	0	41,2	46,0	7	y ₃

5° Scenario Missing At Random (3)

```
PROC MIXED DATA = "Database" METHOD = REML;  
  
CLASS ID new_T;  
  
MODEL new_y = y0(Variabile) new_T;  
  
ESTIMATE 'Treatment Difference At Final Visit'  
new_T 0 0 0 0 0 0 -1 0 0 0 0 0 +1 / CL;  
  
REPEATED / SUBJECT = ID TYPE = UN;  
  
QUIT;
```



Grazie
per
l'attenzione...

