

TECNICHE STATISTICHE PER L'ANALISI DEGLI STUDI LONGITUDINALI

Giorgio Bedogni

INTRODUZIONE AL CORSO

Background 1

“In discussing the analysis of longitudinal data with other researchers, we occasionally meet up with otherwise intelligent researchers who prefer to use simple statistical methods to analyze fairly complex designs.

Having spent thousands if not million of dollars to collect a complex data set, they want to use a simple and cheap statistical tool to analyze their data”

Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005, p. 85.

Background 2

“Medicine has a less strong quantitative tradition and there we occasionally find problems of prejudice against advanced statistical techniques.

Also, lack of proper marketing on the part of the statisticians has undoubtedly contributed to this prejudice”

Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005, p. 100 (corsivo mio).

Background 3

“The process of adapting data to familiar statistical tools is so ubiquitous that has a name: the *shoehorn*.

We *shoehorn* our data into the shoe we have rather than getting a larger shoe as our data grows”

Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005, p. 113.

Calzascarpe? No, grazie!

medidata
Primi negli Studi Osservazionali

Aggiungi agli Studi Clinici
tutte le sfumature
della **Ricerca Clinica Osservazionale**

home
english

AZIENDA | SERVIZI | E-CRF | PROGETTI | SCALE & STRUMENTI PSICOMETRICI | BIBLIOGRAFIA | LINKS | LAVORA CON NOI | FORMAZIONE

Pharma
Biomedicale
NEWS

Benvenuti

MediData srl, è la prima Contract Research Organization (CRO) italiana specializzata nel disegno e nella conduzione di **Studi Osservazionali**.

MediData fornisce consulenza, strumenti e servizi per soddisfare le esigenze di chi opera nel mercato farmaceutico e biomedicale, e di chi conduce attività di studio e ricerca medico-scientifica in enti no-profit.

La nostra **mission** è **ESSERE PRIMI NEGLI STUDI OSSERVAZIONALI** grazie alla competenza, esperienza e flessibilità che mettiamo al servizio dei vostri progetti!

Gli Studi Osservazionali

Gli **Studi Osservazionali** e gli studi sugli esiti (outcomes research) offrono l'opportunità di arricchire e completare i risultati della sperimentazione clinica con informazioni strategiche tratte dalla normale pratica medica, con particolare focus sugli outcomes clinici (effectiveness e safety), economici ed umanistici.

Studi Clinici osservazionali in tutte le loro sfumature

Epidemiologia Clinica, Farmacoeconomia, Outcome Research, Patient Reported Outcomes, Disease Management, Monitoraggio Linee Guida, PASS, Registri, Drug Utilization Studies, ...

01/06/2009
Altre news...
Contatta Medidata

MediData s.r.l. c/o Centro Servizi CittàNova - Viale Virgilio 54/U, 41100 MODENA, Italy Tel. 059 88.60.134 059 88.60.135 - Fax 059 84.83.05 p.iva 02041030350 - Medidata mail - powered by EXPERTWEB

Obiettivi

- Passare in rassegna le principali tecniche statistiche per l'analisi longitudinale
- Fornire indicazioni utili per la loro applicazione agli studi clinici

Requisiti

- Familiariatà con l'analisi della varianza e la regressione lineare

Strategia didattica 1

- Il corso è incentrato sugli aspetti pratici dell'analisi longitudinale
- Pertanto
 - Ho presentato il minor numero possibile di formule
 - Ho scelto una notazione statistica semplificata

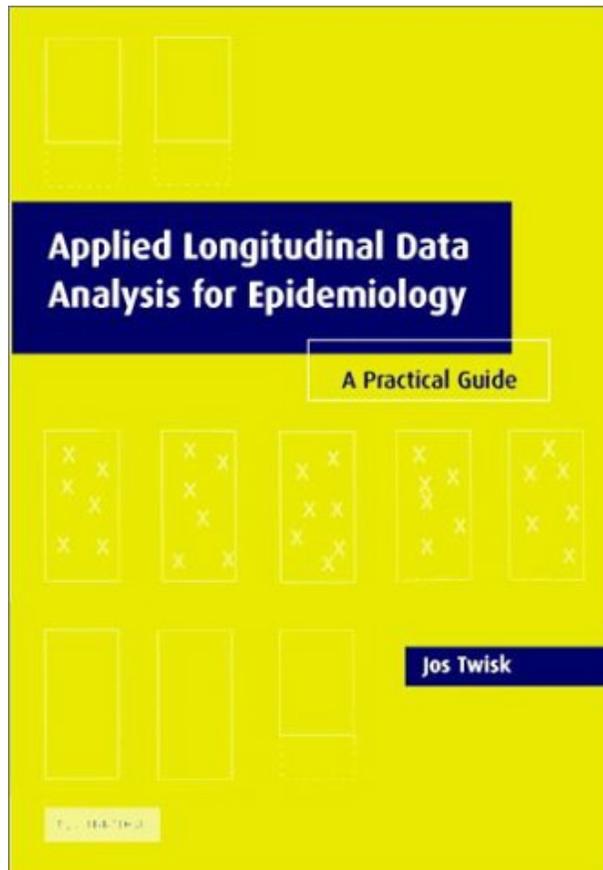
Strategia didattica 2

- 12 (4+4+4) ore di corso su un argomento così vasto rischiano di violare un principio sacrosanto:

"Make everything as simple as possible, but not simpler" (Albert Einstein)

ma correrò il rischio...

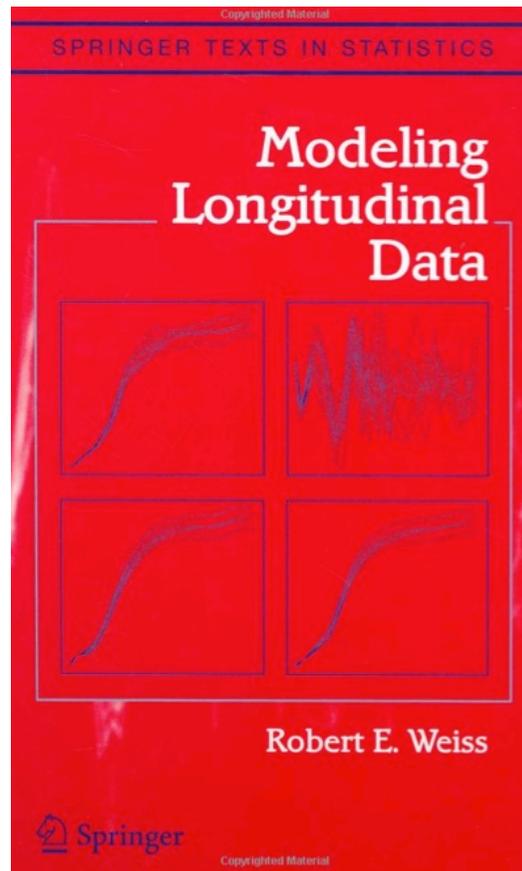
Testo consigliato 1



Twisk JWR. *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge: Cambridge University Press, 2003.

Ottima introduzione all'analisi longitudinale - livello principiante/intermedio

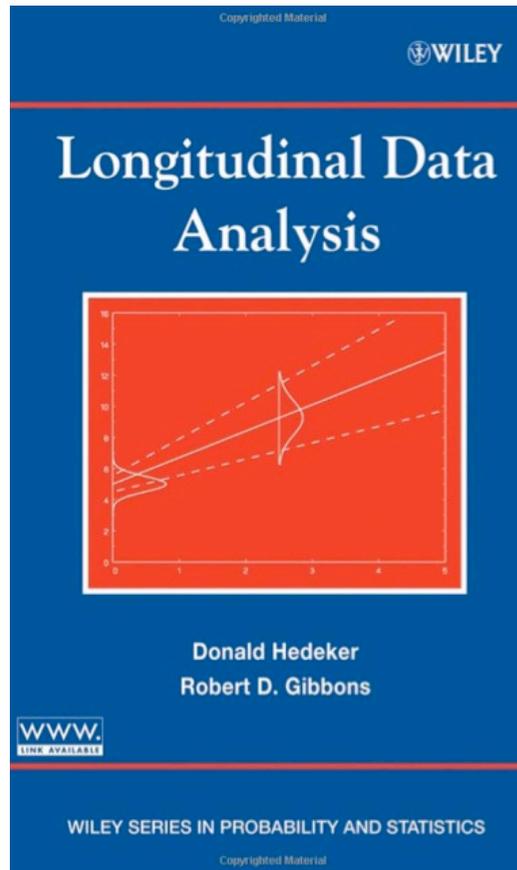
Testo consigliato 2



Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005.

Ottimo compromesso tra la teoria e la pratica - livello intermedio/avanzato

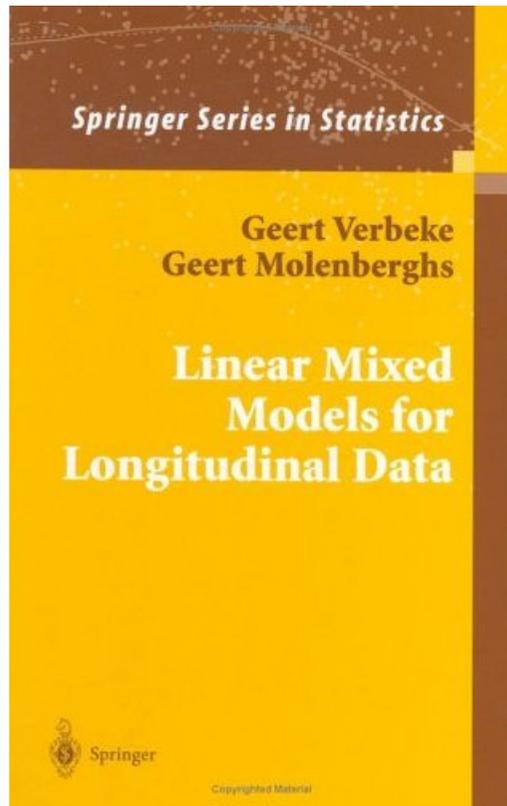
Testo consigliato 3



Hedeker DR, Gibbons RD.
Longitudinal data analysis. New
York: Wiley, 2006.

Buona trattazione "tradizionale"
- livello intermedio/avanzato

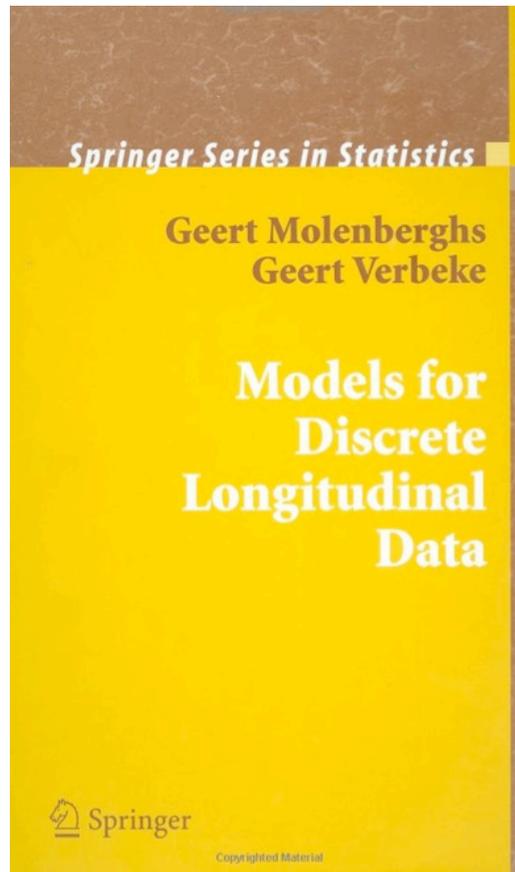
Testo consigliato 4



Verbeke G, Molenberghs G.
*Linear mixed models for
longitudinal data*. New York:
Springer, 2000.

Buona trattazione tecnica con
esempi pratici - livello avanzato

Testo consigliato 5



Molenberghs G, Verbeke G.
Models for discrete longitudinal data. New York: Springer,
2005.

Buona trattazione tecnica con
esempi pratici - livello avanzato

Testo consigliato 6

ANALYSIS OF LONGITUDINAL STUDIES IN
EPIDEMIOLOGY

NICHOLAS P. JEWELL & ALAN HUBBARD¹

January 18, 2006

Jewell NP, Hubbard A. *Analysis of longitudinal studies in epidemiology*. London: Chapman & Hall, 2010 (in press).

Early draft disponibile su:
<http://ehs.sph.berkeley.edu/hubbard/longdata/>

Ottima trattazione con esempi pratici - livello intermedio/avanzato

¹Send correspondence to Nicholas P. Jewell, Division of Biostatistics, School of Public Health, 140 Warren Hall #7360, University of California, Berkeley, CA 94720, USA. Tel: 510-642-4627, Fax: 510-643-5163, e-mail: jewell@stat.berkeley.edu ©Nicholas P. Jewell & Alan Hubbard

Materiale di supporto

- Dispensa del corso (PDF stampato)
- Articoli (PDF elettronici)
- www.giorgiobedogni.it/tutorials/ptutorials.html

Programma

- Modulo 1 (10/07/09)
- Modulo 2 (24/07/09)
- Modulo 3 (31/07/09)

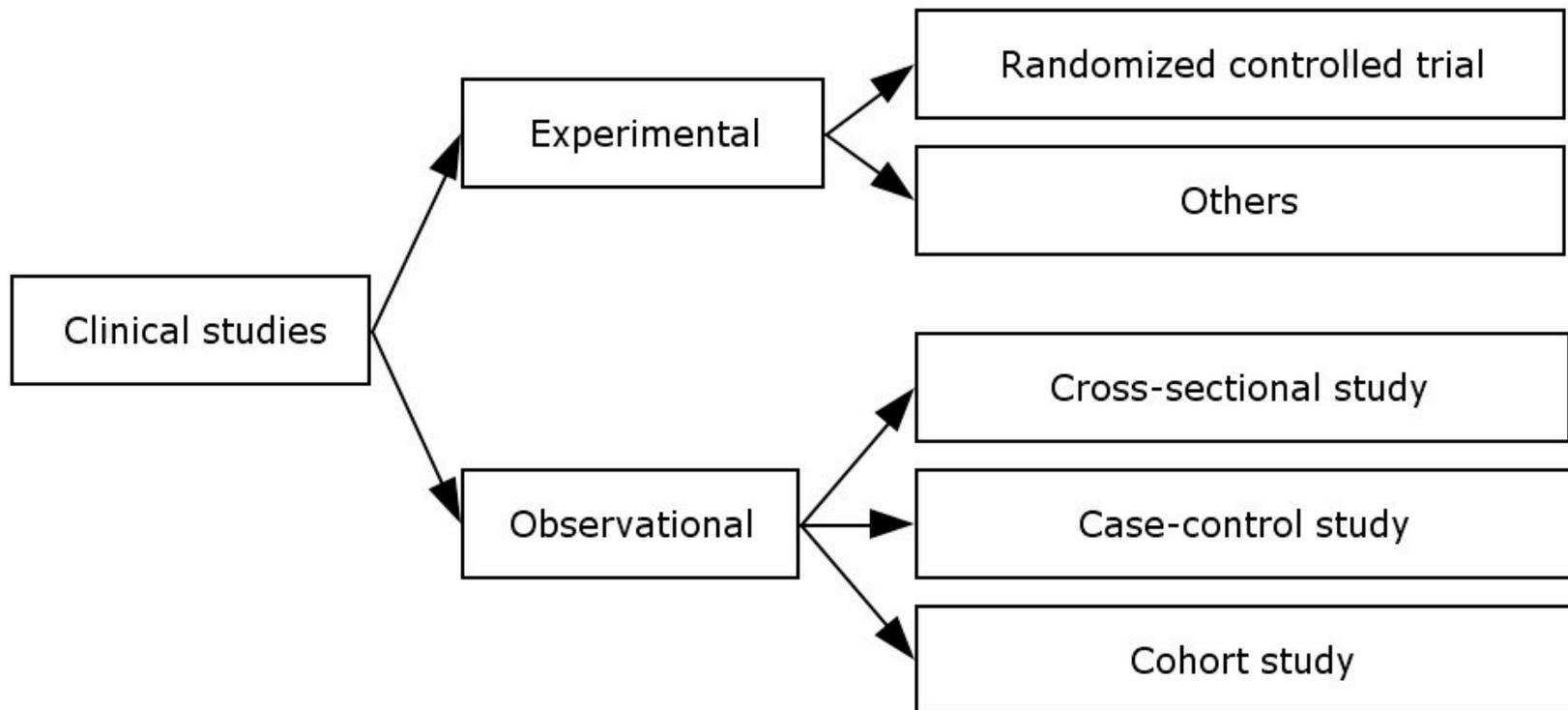
MODULO 1

Modulo 1

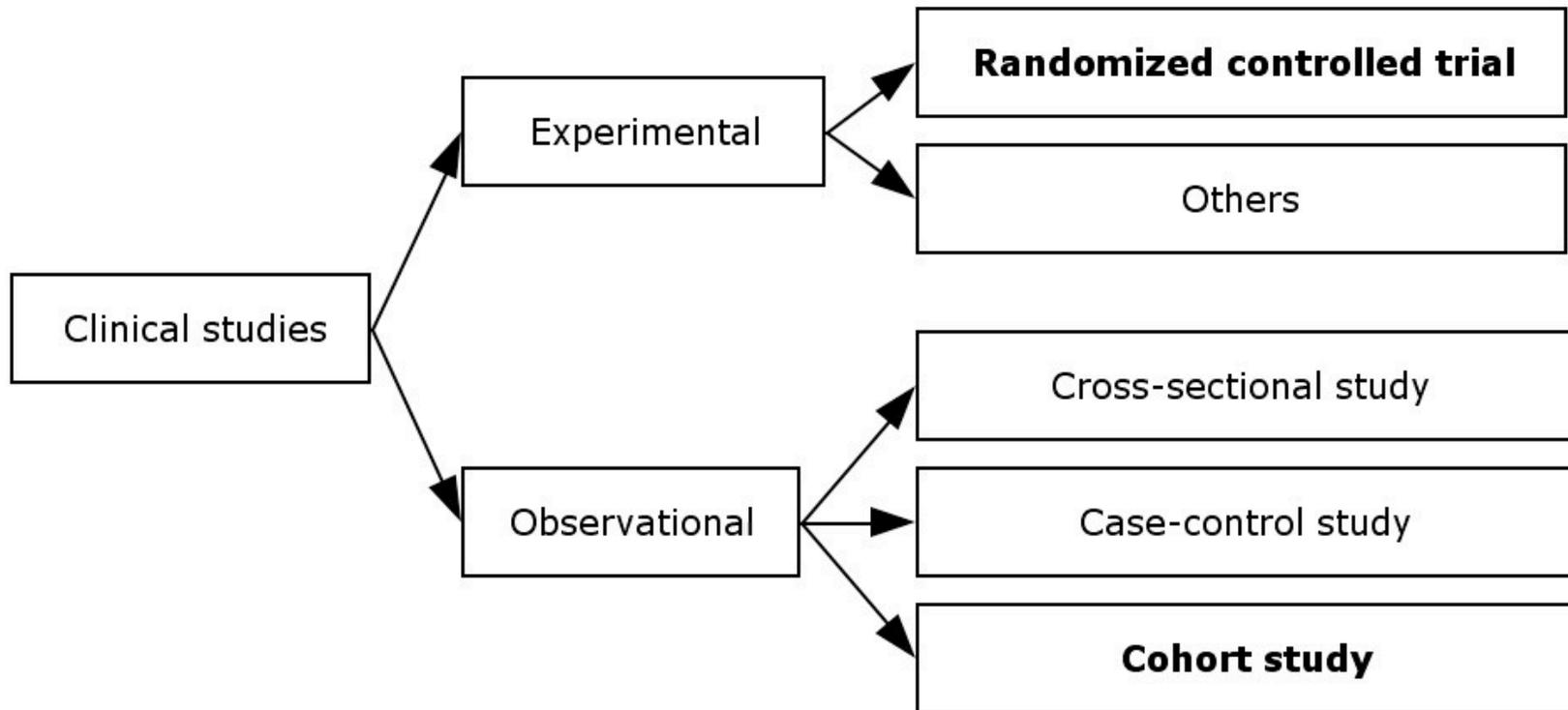
- Introduzione agli studi longitudinali
- Descrizione dei dati longitudinali
- Analisi dei dati longitudinali: sinossi
- Analisi di outcome continui 1
 - *NB: da questo punto in poi il corso è per statistici*

INTRODUZIONE AGLI STUDI LONGITUDINALI

Studi clinici



Studi longitudinali



Cosa è uno studio longitudinale?

- Uno studio la cui variabile di *outcome* è misurata più di una volta
- (Ricordate la distinzione tra variabili di *outcome* e predittori che abbiamo fatto al Corso sulla Metodologia degli Studi Osservazionali?)

Uno studio longitudinale in tre aggettivi

- Multivariato
- Gerarchico o multilivello
- Ordinato nel tempo

Cosa significa multivariato?

- Multivariato ("multivariate")
 - Più di un outcome per individuo
 - ≠ multivariabile ("multivariable")!

Multivariato non è multivariabile

	Y_1	Y_1, Y_2, \dots, Y_n
X_1	Univariato Univariabile	Multivariato Univariabile
X_1, X_2, \dots, X_n	Univariato Multivariabile	Multivariato Multivaribile

Cosa significa “gerarchico” o “multilivello”?

- Gerarchico o multi-livello (“hierarchical” or “multilevel”)
 - Le osservazioni sono clusterizzate (“clustered”) o nidificate (“nested”) per soggetto
 - Il tempo è definito livello 1 (“level 1”) e l’individuo livello 2 (“Level 2”)
 - *Il livello 2 contiene il livello 1 (e non viceversa)*

Cosa significa "ordinato nel tempo"?

- Ordinato nel tempo
 - Le misure ripetute sono "ordinate" secondo il tempo
 - Riuscite a immaginare misure ripetute non ordinate secondo un criterio temporale?

Qual è la caratteristica distintiva di uno studio longitudinale?

- Le osservazioni ripetute dello stesso individuo sono (per lo più) positivamente correlate
- I metodi statistici standard assumono osservazioni *independenti*
- Utilizzare i metodi statistici standard per analizzare dati longitudinali comporta (per lo più) una sottostima dell'errore standard (=esagerazione dell'effetto) e del valore di p (=esagerazione della significatività statistica)

Quali sono i vantaggi di uno studio longitudinale 1?

- Il vantaggio principale è la possibilità di studiare lo “sviluppo individuale” o “individual development” di un determinato outcome
- Tale sviluppo può essere messo in relazione con lo sviluppo di uno più predittori

Quali sono i vantaggi di uno studio longitudinale 2?

- Un altro vantaggio è che uno studio longitudinale soddisfa il criterio di temporalità per la valutazione di un rapporto causa-effetto (la causa deve precedere l'effetto)
- (Ricordate la discussione critica della "causalità" che abbiamo fatto al Corso sulla Metodologia degli Studi Osservazionali?)

Quali sono i limiti di uno studio longitudinale?

- Dati mancanti o “missing data”

Chi sono i protagonisti di uno studio longitudinale?

- Individuo
 - $i =$ da 1 a N
- Tempo
 - $t =$ da 1 a T
- Outcome
 - Y
- Predittore/ i
 - X

Come si presentano i dati?

"Long" vs. "wide" format

id	iauc	year
1	2446.438	1
1	8144.869	2
1	7837.075	3
1	5004.608	4
1	5941.062	5
2	861.4052	1
2	2548.252	2
2	1916.569	3
2	1192.141	4
2	734.3137	5
3	4687.614	1
3	3310.311	2
3	3400.278	3

iauc1	iauc2	iauc3	iauc4	iauc5
2446.438	8144.869	7837.075	5004.608	5941.062
861.4052	2548.252	1916.569	1192.141	734.3137
4687.614	3310.311	3400.278	.	.

Come si presenta il tempo? "Equally spaced" vs. "unequally spaced"

id	iauc	year
1	2446.438	1
1	8144.869	2
1	7837.075	3
1	5004.608	4
1	5941.062	5
2	861.4052	1
2	2548.252	2
2	1916.569	3
2	1192.141	4
2	734.3137	5
3	4687.614	1
3	3310.311	2
3	3400.278	3

id	day	cd4
1	39	45
1	137	119
1	147	113
1	179	74
1	187	95
1	214	120
1	256	209
1	274	375
1	298	137
1	335	.
1	354	167
1	411	.
1	438	198
1	508	157
1	573	220
1	634	239
1	691	411
1	760	262
1	802	.
1	837	397
1	956	465
1	1076	426
1	1159	423
1	1173	.
1	1264	398
1	1355	433

id	day	cd4
2	0	196
2	7	369
2	13	353
2	27	474
2	55	425
2	111	493
2	139	464
2	167	448
2	195	427
2	223	460
2	251	484
2	279	513
2	335	467
2	419	618
2	503	487
2	587	777
2	671	650
2	756	575
2	839	723
2	923	617
2	1007	594
2	1091	675
2	1204	750
2	1259	.
2	1314	.
2	1343	.
2	1441	.
2	1442	664

DESCRIZIONE DEI DATI LONGITUDINALI

Quali metodi?

- Statistiche riassuntive
 - Standard
 - Evolute
- Metodi grafici

Statistiche standard

- Media e deviazione standard
 - Richiedono distribuzione gaussiana
- Mediana e percentili
 - Applicabili a distribuzioni non gaussiane
 - (Richiedono variabile "trattabile" come continua)

Statistiche evolute

Analysis of serial measurements in medical research

J N S Matthews, Douglas G Altman, M J Campbell, Patrick Royston

BMJ VOLUME 300 27 JANUARY 1990

Statistiche evolute

Type of data	Property	Measure
Peaked	Overall value of response	Mean or area under the curve (AUC)
Peaked	Value of most extreme response	Maximum (minimum)
Peaked	Delay in response	Time to maximum or minimum
Growth	Rate of change of response	Linear regression coefficient
Growth	Final level of response	Final value or relative difference between first and last
Growth	Delay in response	Time to reach a particular value

Matthews JNS *et al.* *BMJ* 1990, **300**:230.

Statistiche riassuntive evolute

- Attenzione: sono (e le presentiamo come) statistiche riassuntive
- Utilizzarle per l'analisi longitudinale vera e propria è appropriato solo in pochi contesti

Metodi grafici

“There is no single statistical tool that is powerful as a well-chosen graph”

“Eye-balling can give diagnostic insights no formal diagnostic will ever provide”

Chatterjee S, Hadi AS. *Regression analysis by example*. New York: Wiley, 2006.

Metodi grafici

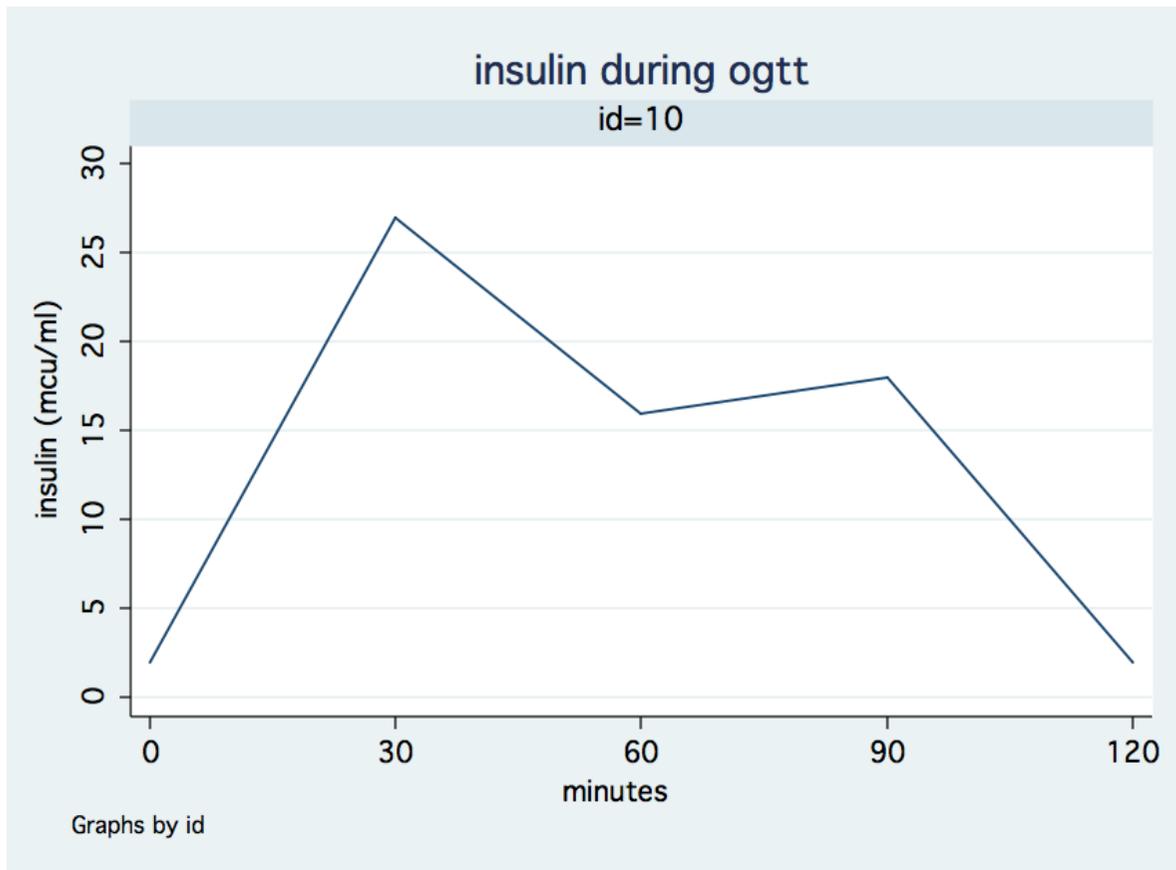
- Essenziali per rispondere alla domanda: sono presenti associazioni *biologicamente o clinicamente* rilevanti?
- Cominciare sempre studiando la distribuzione dell' outcome (quello che faremo ora)
- Tenere bene a mente che i grafici sono funzione della domanda dello studio e non esistono grafici "universali"

Case-study

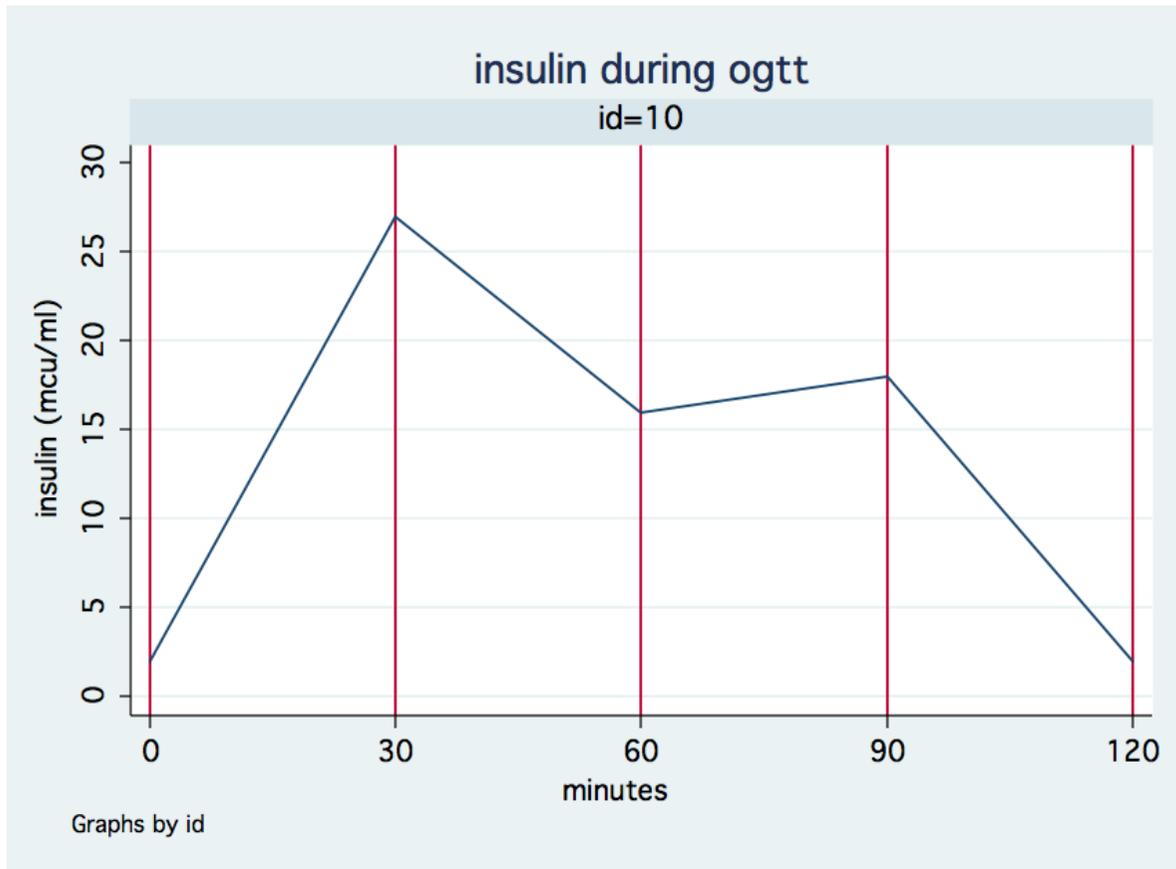
- Studio di coorte della durata di 5 anni
- 37 bambini HIV+ in trattamento con inibitori delle proteasi
- Misurazioni annuali dell'area sotto la curva dell'insulinemia (iauc) durante carico orale di glucosio (OGTT)

Viganò A *et al.* *Clin Drug Investig* 2009;**29**:101.

Come calcolo Iniuac?



Come calcolo Iniuac? ("trapezoidal rule")

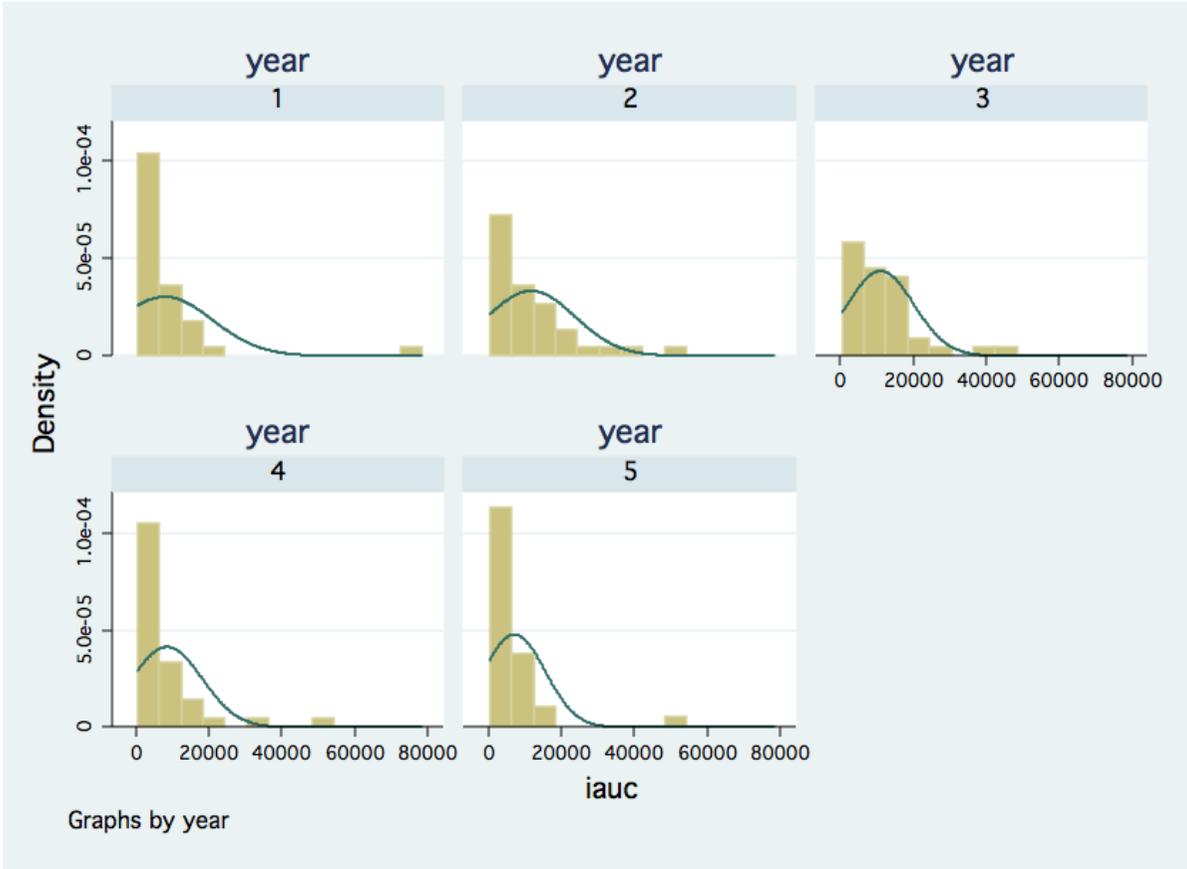


Statistica riassuntiva

Summary for variables: iauc
by categories of: year

year	N	mean	sd	min	max
1	37	7915	13140	426	78379
2	37	11793	11971	646	53169
3	37	11036	9154	1866	42958
4	35	8523	9653	901	50040
5	31	7199	8363	734	48609
Total	177	9373	10720	426	78379

Normalità?



Normalità?

-> year==1

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
iauc	37	0.50082	18.587	6.121	0.00000

-> year==2

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
iauc	37	0.81692	6.817	4.020	0.00003

-> year==3

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
iauc	37	0.80890	7.116	4.110	0.00002

-> year==4

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
iauc	35	0.64802	12.563	5.283	0.00000

-> year==5

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
iauc	31	0.52872	15.351	5.659	0.00000

Statistica riassuntiva

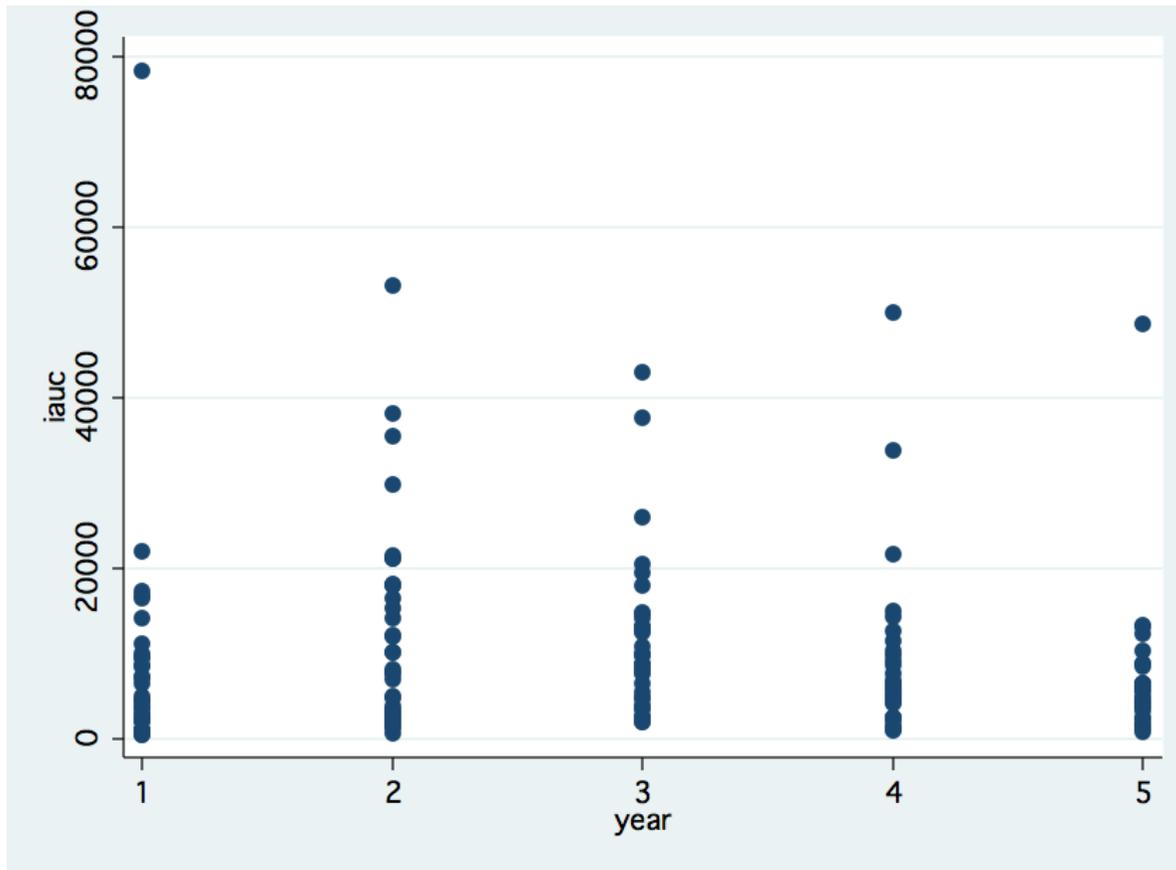
Summary for variables: iauc
by categories of: year

year	N	p50	iqr	min	max
1	37	4247	7594	426	78379
2	37	7830	15182	646	53169
3	37	8828	8293	1866	42958
4	35	5809	5740	901	50040
5	31	5830	4863	734	48609
Total	177	5990	9158	426	78379

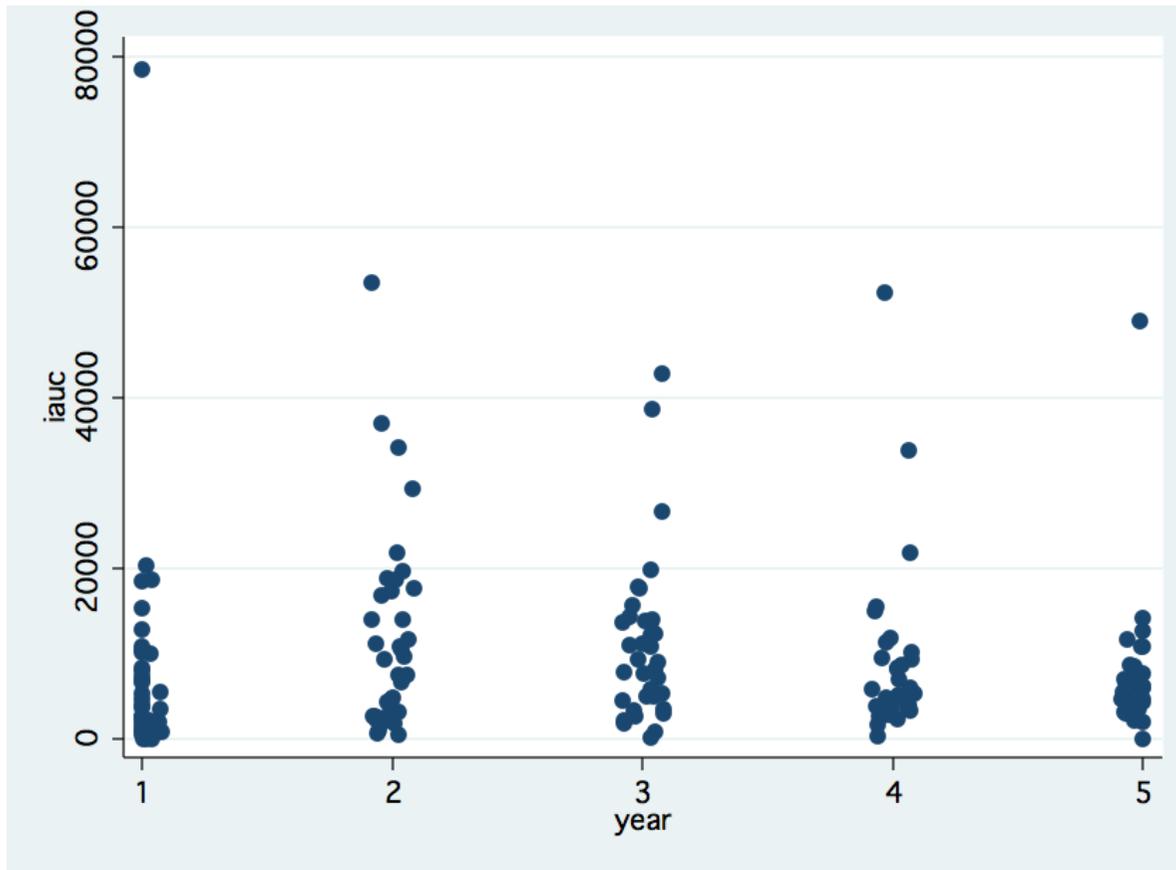
Problemi delle statistiche riassuntive

- I dati ai differenti tempi non sono indipendenti
- Il numero di individui diminuisce all'aumentare del follow-up

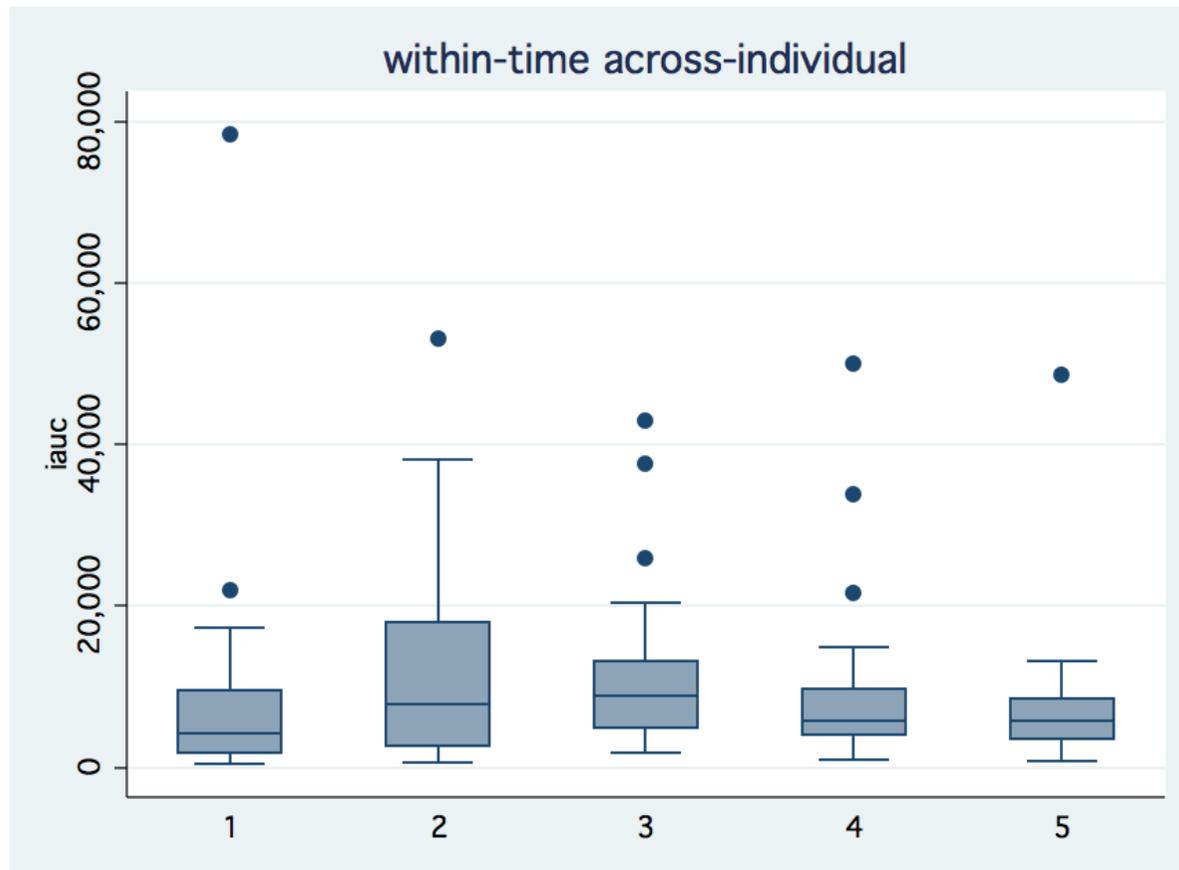
Scatterplot semplice



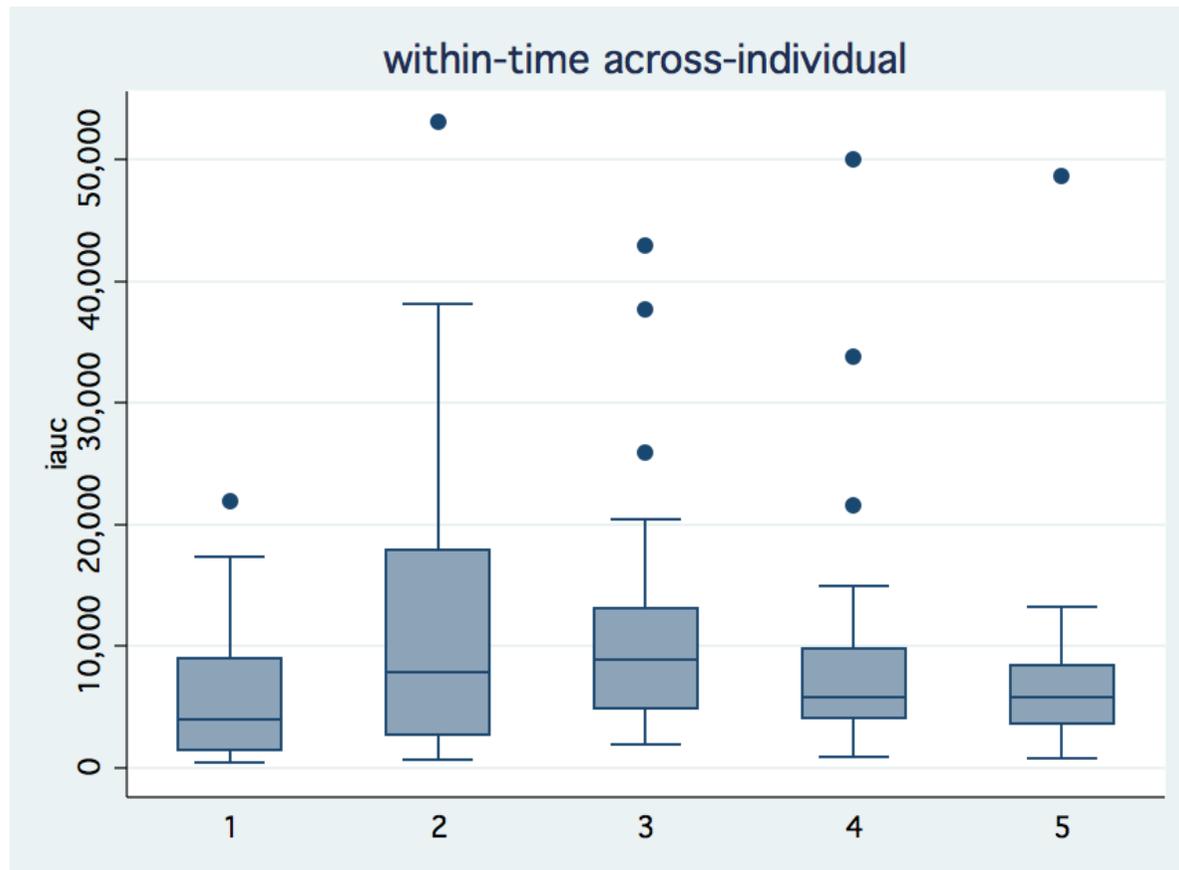
Scatterplot con jitter



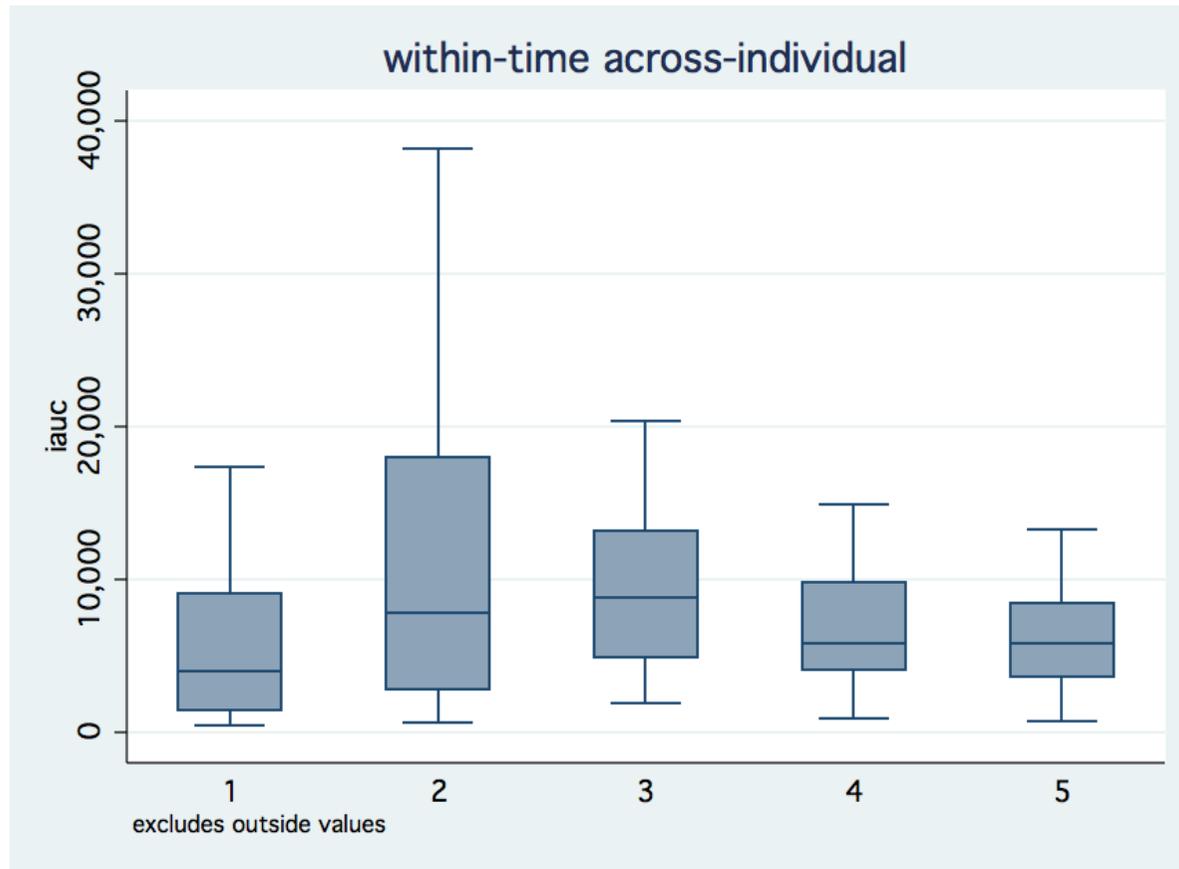
Boxplot



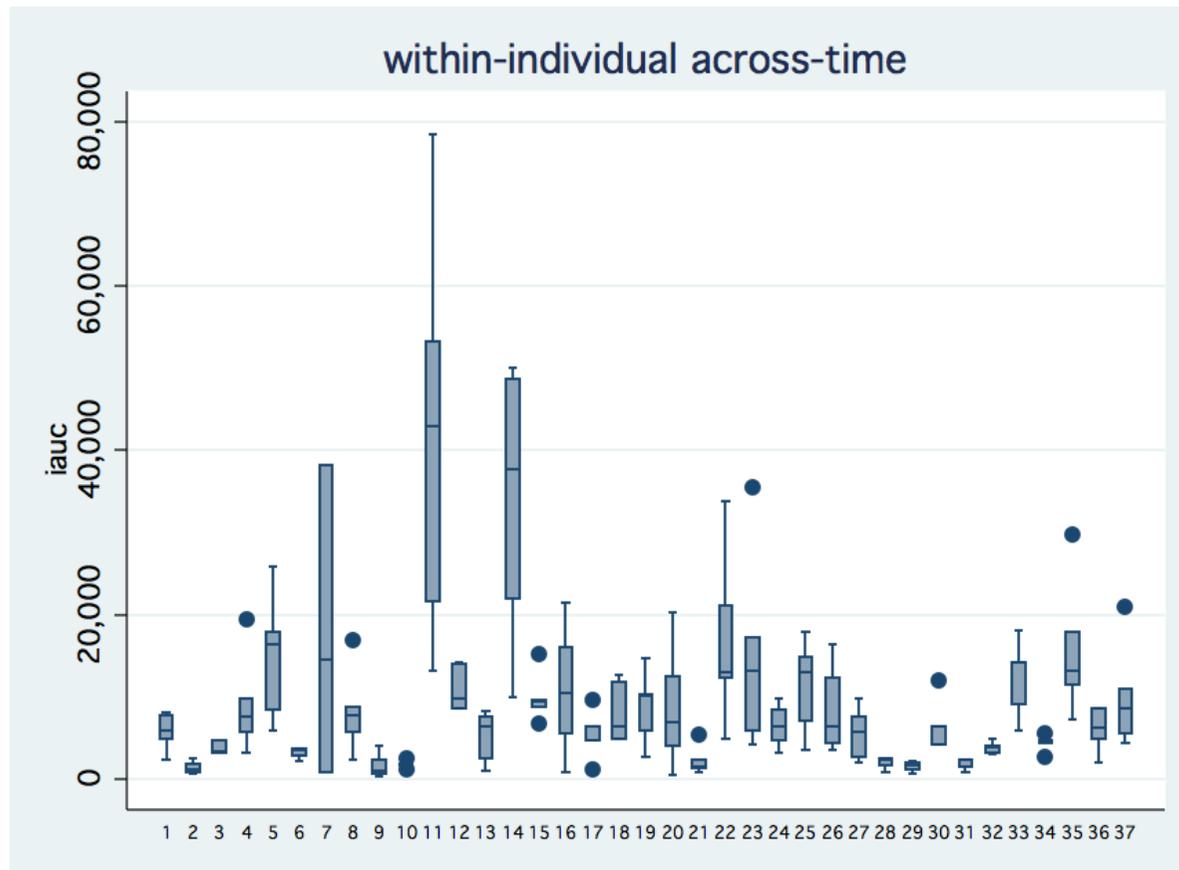
Boxplot con esclusione di "outlier"



Boxplot senza "outsiders" (mai in prima battuta)



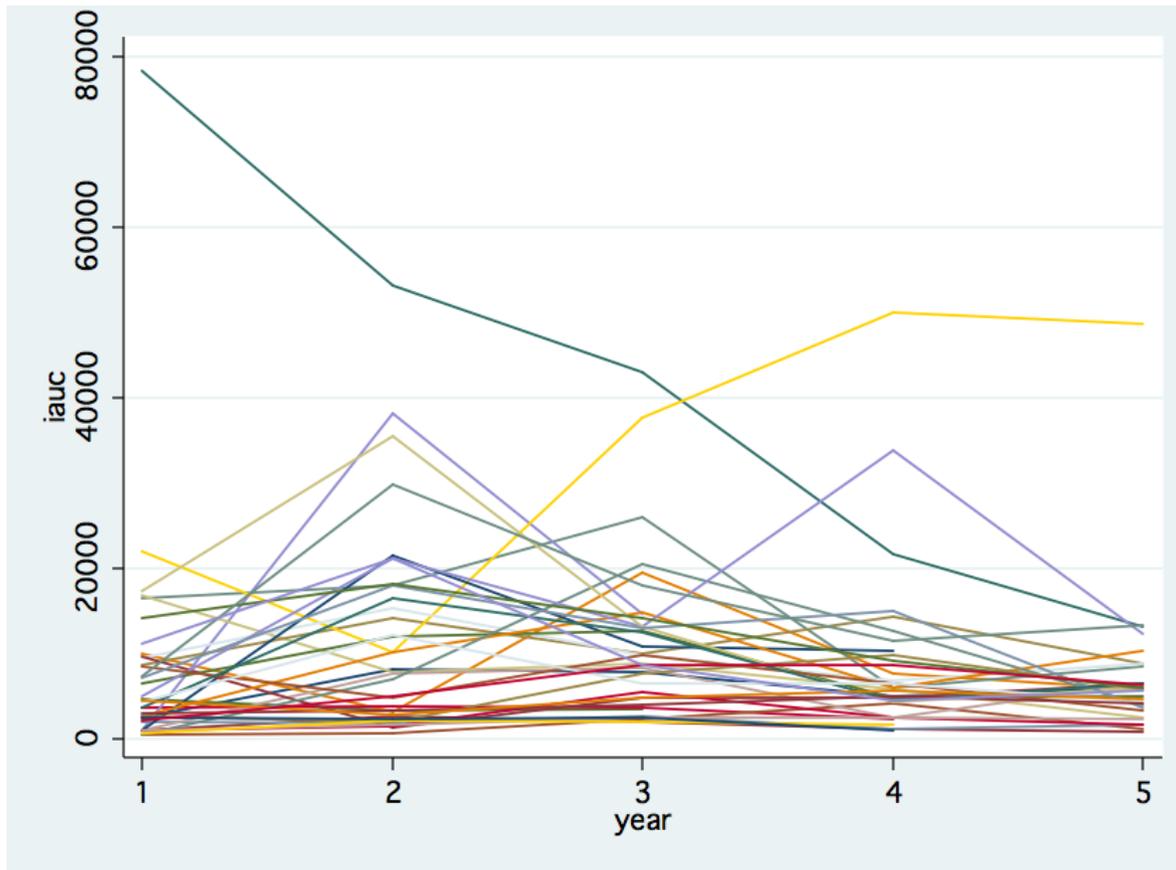
Non dimenticate di visualizzare la variabilità
intra-individuale !



Problemi di scatterplots e boxplots

- I dati ai differenti tempi non sono indipendenti
- Il numero di individui diminuisce all'aumentare del follow-up
- (Gli stessi delle statistiche riassuntive)

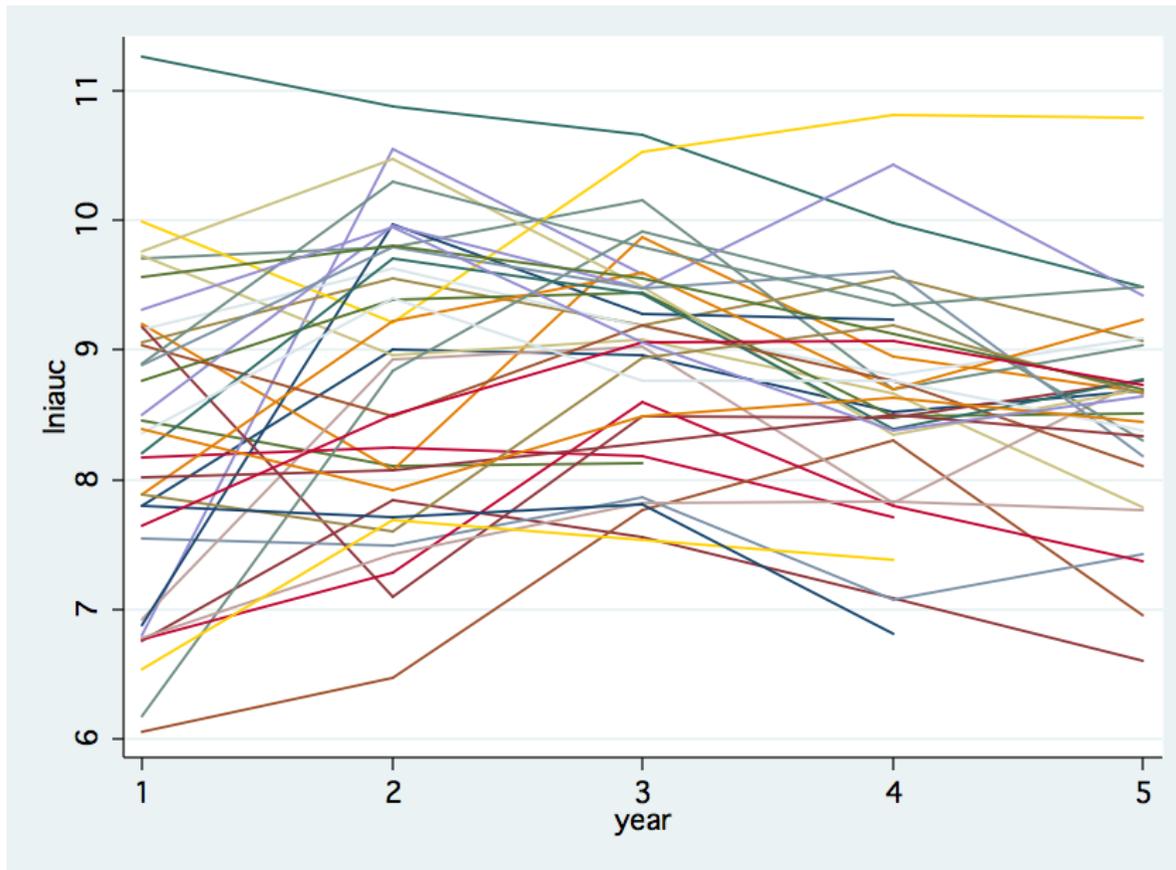
Profileplot



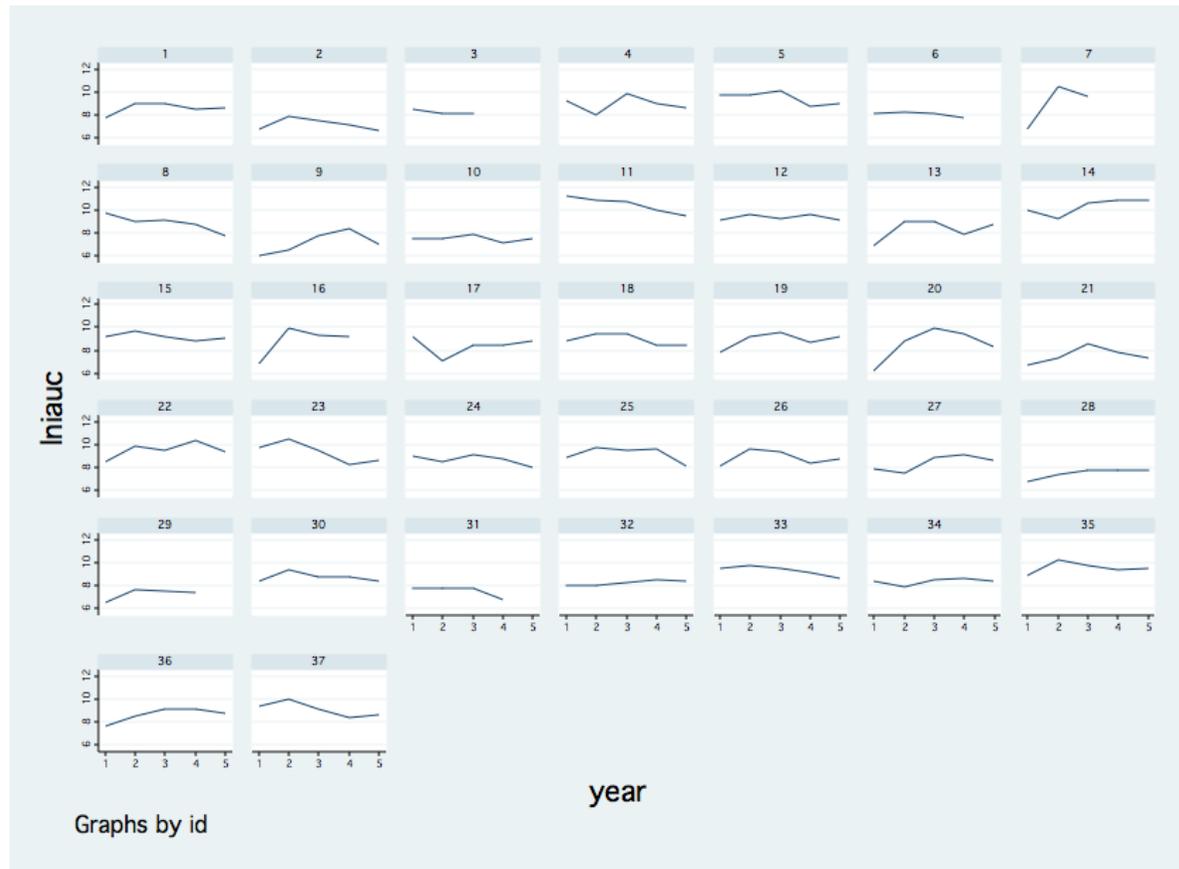
Profileplot con trasformazione di Y

- Ho utilizzato il logaritmo naturale (= base e) di Y
 - $\ln_{iauc} = \ln(iauc)$

Profileplot con trasformazione di Y



Profileplots individuali o "trellis plots" con trasformazione di Y



Profileplot con trasformazione di Y

- Se ci sono molti individui e/o tempi, qualsiasi trasformazione è inefficiente perché la densità dei punti è tale da impedire di cogliere i profile-plots (vedi prossimo case-study)

Come visualizzo la "rate of change"?

Type of data	Property	Measure
Peaked	Overall value of response	Mean or area under the curve (AUC)
Peaked	Value of most extreme response	Maximum (minimum)
Peaked	Delay in response	Time to maximum or minimum
Growth	Rate of change of response	Linear regression coefficient
Growth	Final level of response	Final value or relative difference between first and last
Growth	Delay in response	Time to reach a particular value

Come visualizzo la "rate of change"?

- La velocità di cambiamento è il coefficiente angolare (= slope) della retta ottenuta dalla regressione lineare di Y (lniauc) su t (year)

Caveat

- Stiamo assumendo una relazione lineare tra l'outcome e il tempo (e ogni assunzione va verificata...)
- Ci sono "missing data" e la regressione non utilizza gli stessi punti per tutti i soggetti

Calcolo dello slope

id	slope	inter	df
1	.128748	8.210138	3
2	-.1078914	7.492167	3
3	-.160533	8.550762	1
4	-.0177997	9.006038	3
5	-.2431521	10.20993	3
6	-.1450467	8.444845	2

id = identificatore del paziente,
da 1 a 37

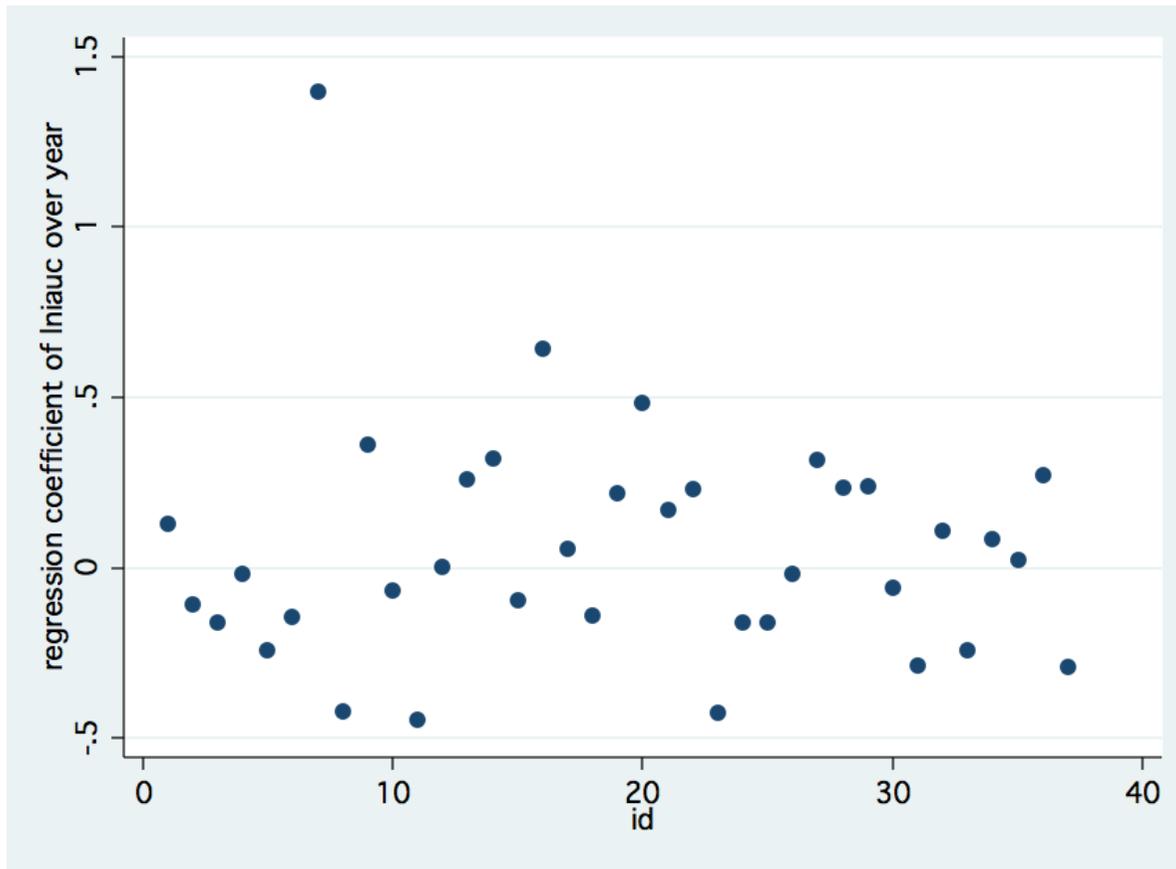
slope = coefficiente di
regressione

Inter = intercetto

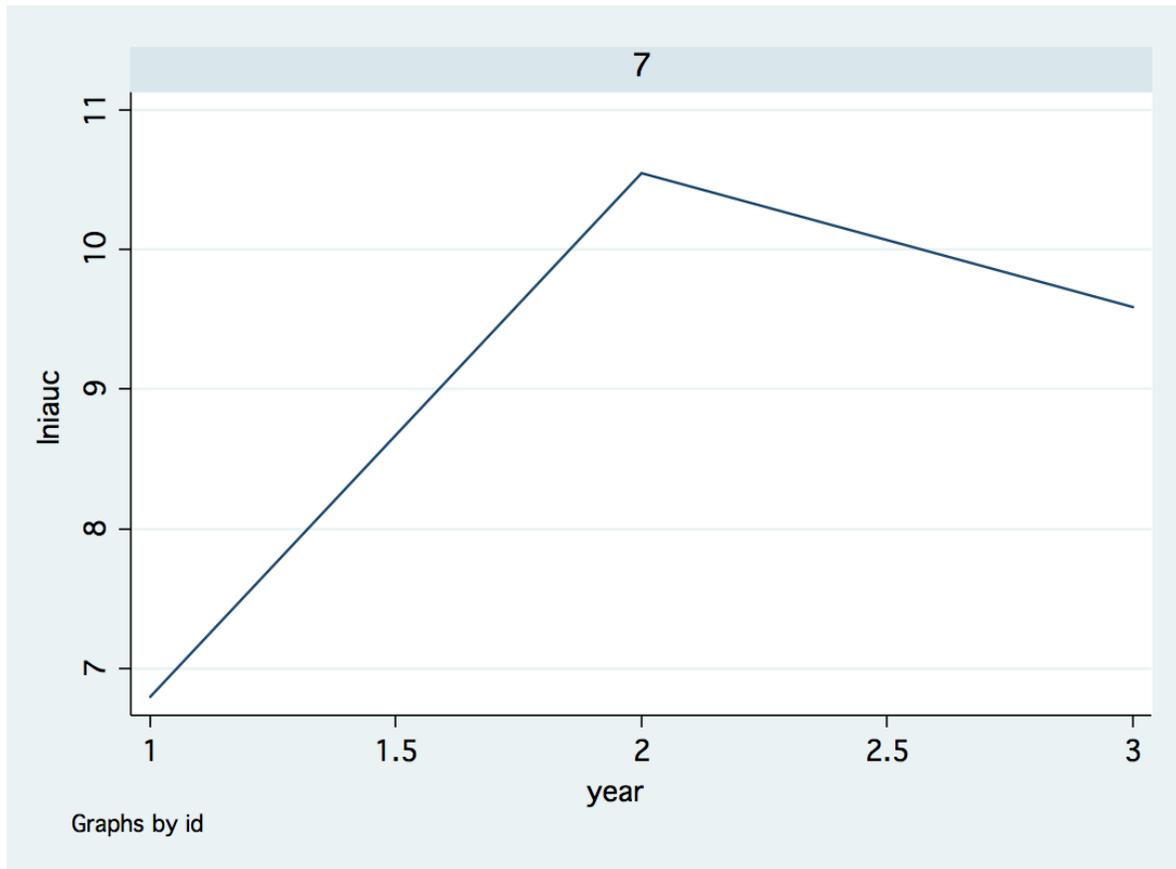
df = gradi di libertà

Perché i gradi di libertà
differiscono da soggetto a
soggetto?

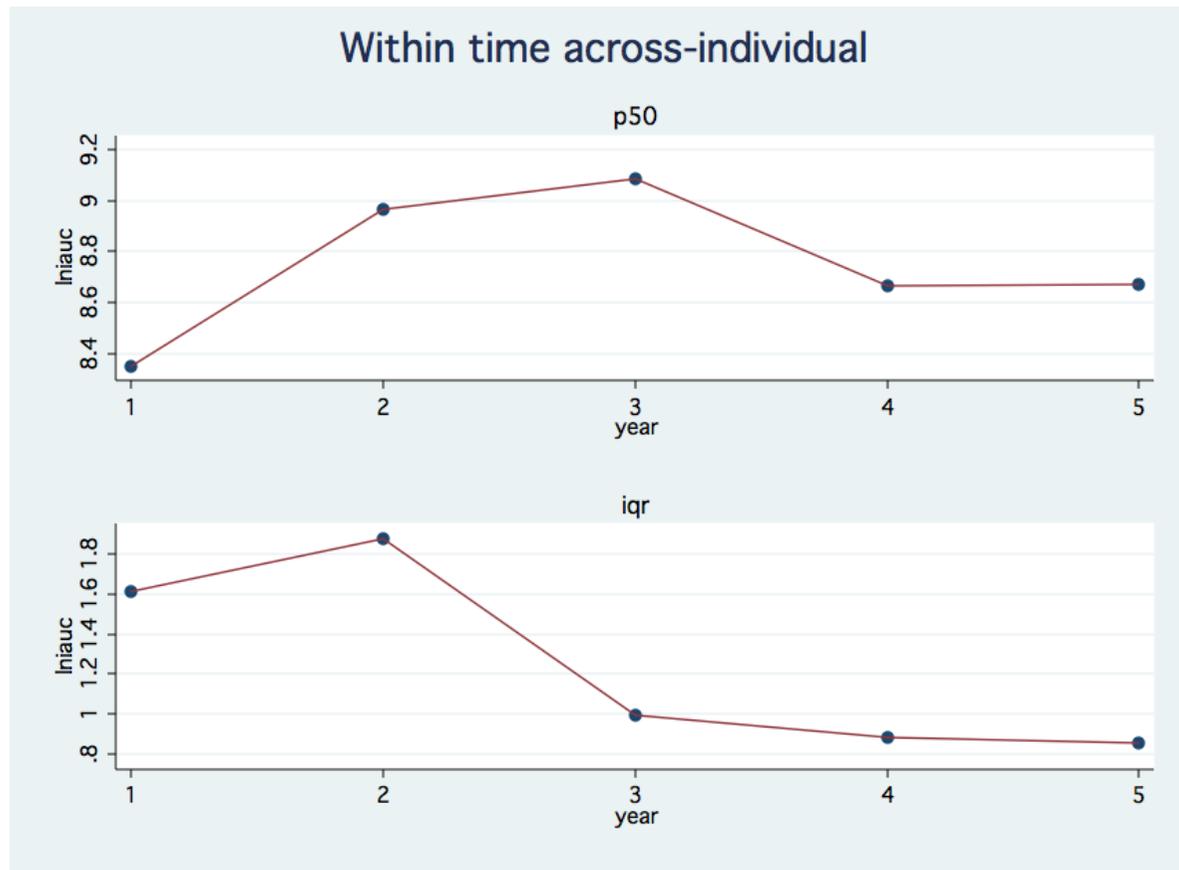
Distribuzione dello slope



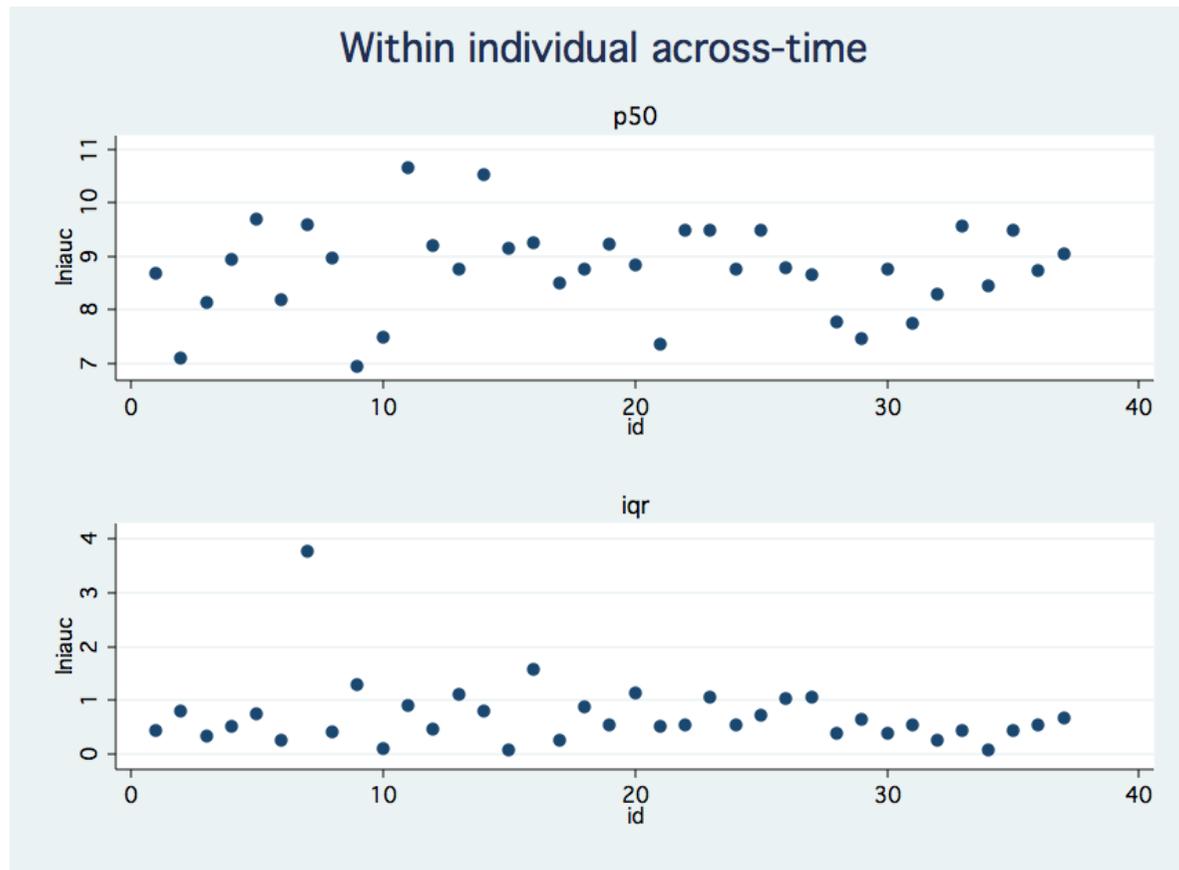
Questo è il soggetto con lo slope = 1.5
Attendo commenti...



Due variazioni sul tema dello scatterplot



Due variazioni sul tema dello scatterplot

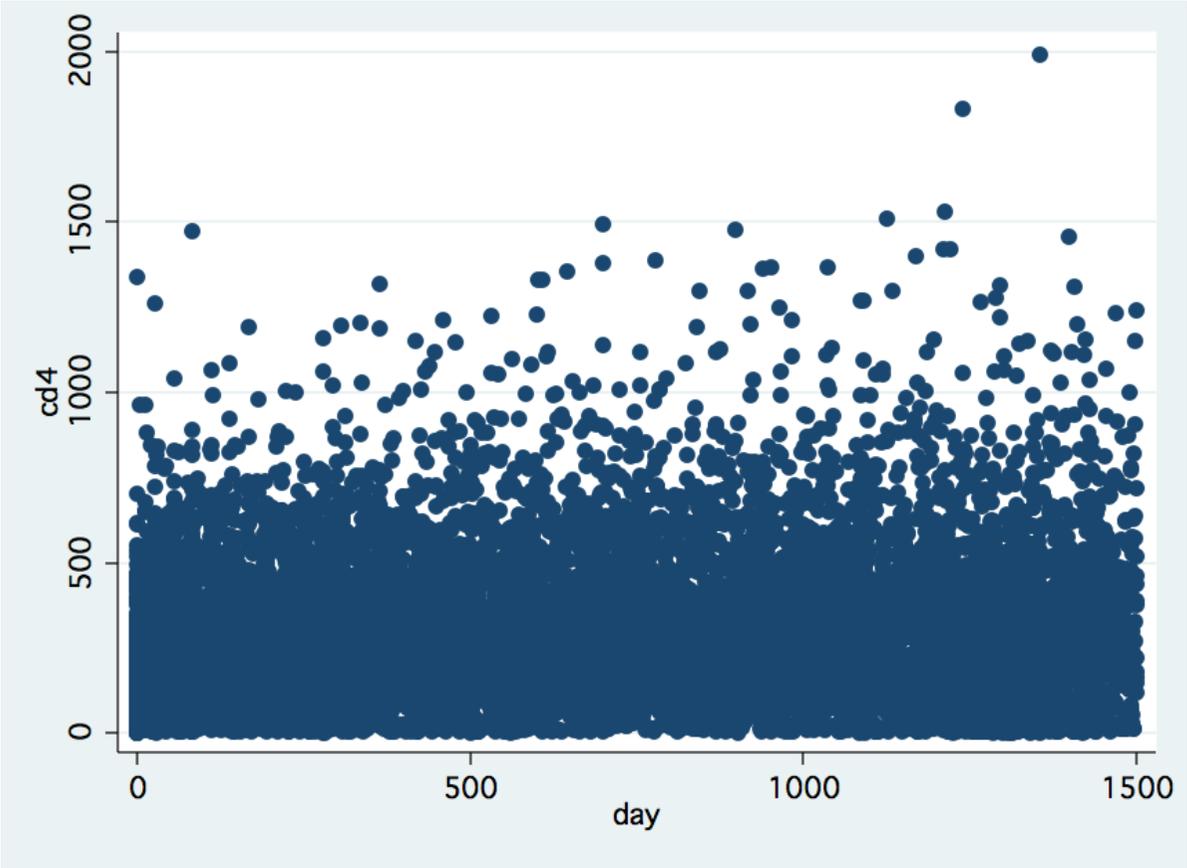


Case-study

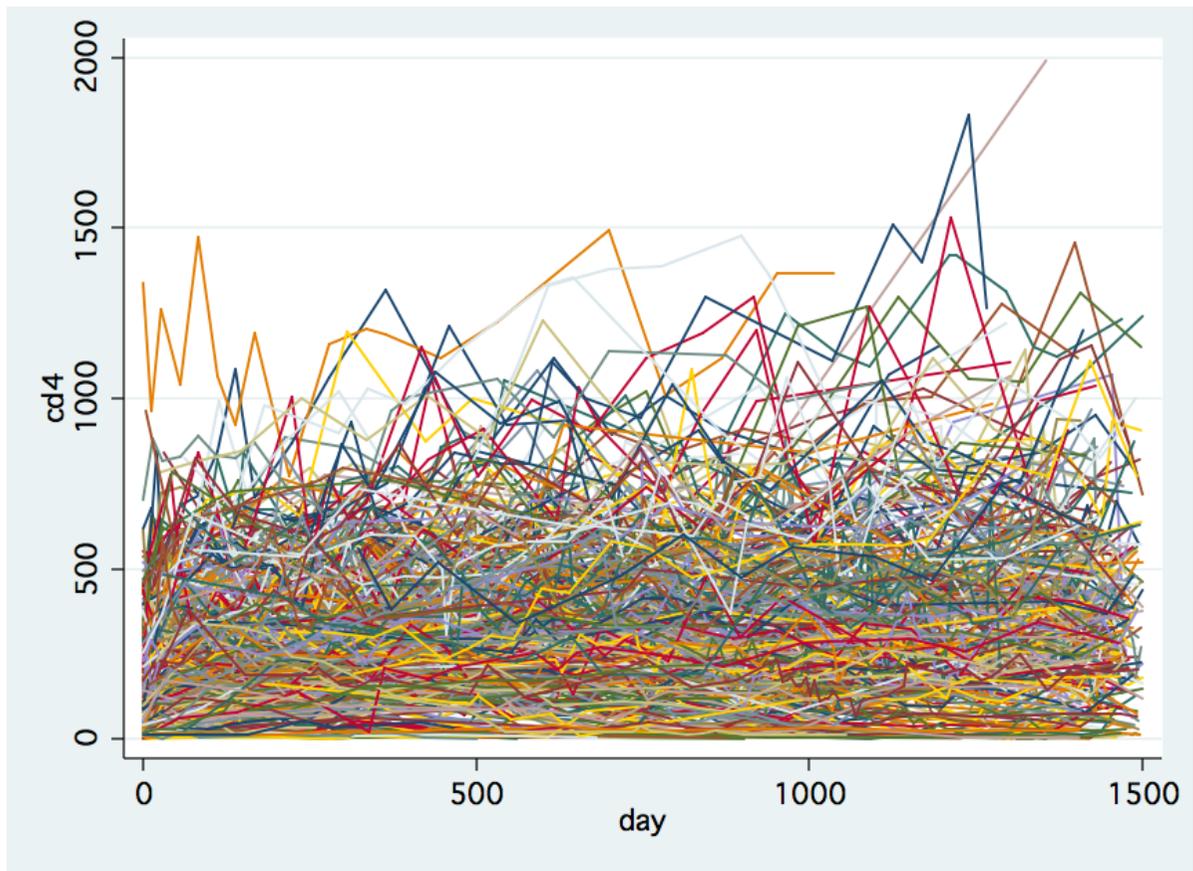
- “Messy data”
- 483 soggetti HIV+ con numero variabile di misurazioni “unequally spaced” (da 11 a 89)
- $Y = \text{cd4} = \text{numero linfociti CD4}$
- $t = \text{days} = \text{giorni}$

<http://ehs.sph.berkeley.edu/hubbard/longdata/> (cd4b.csv)

Scatterplot?



Profileplot?



Trellis plots?

- Di 483 soggetti ? Per di più con un numero di punti variabile da 11 a 89?
- Non sembra una buona idea...

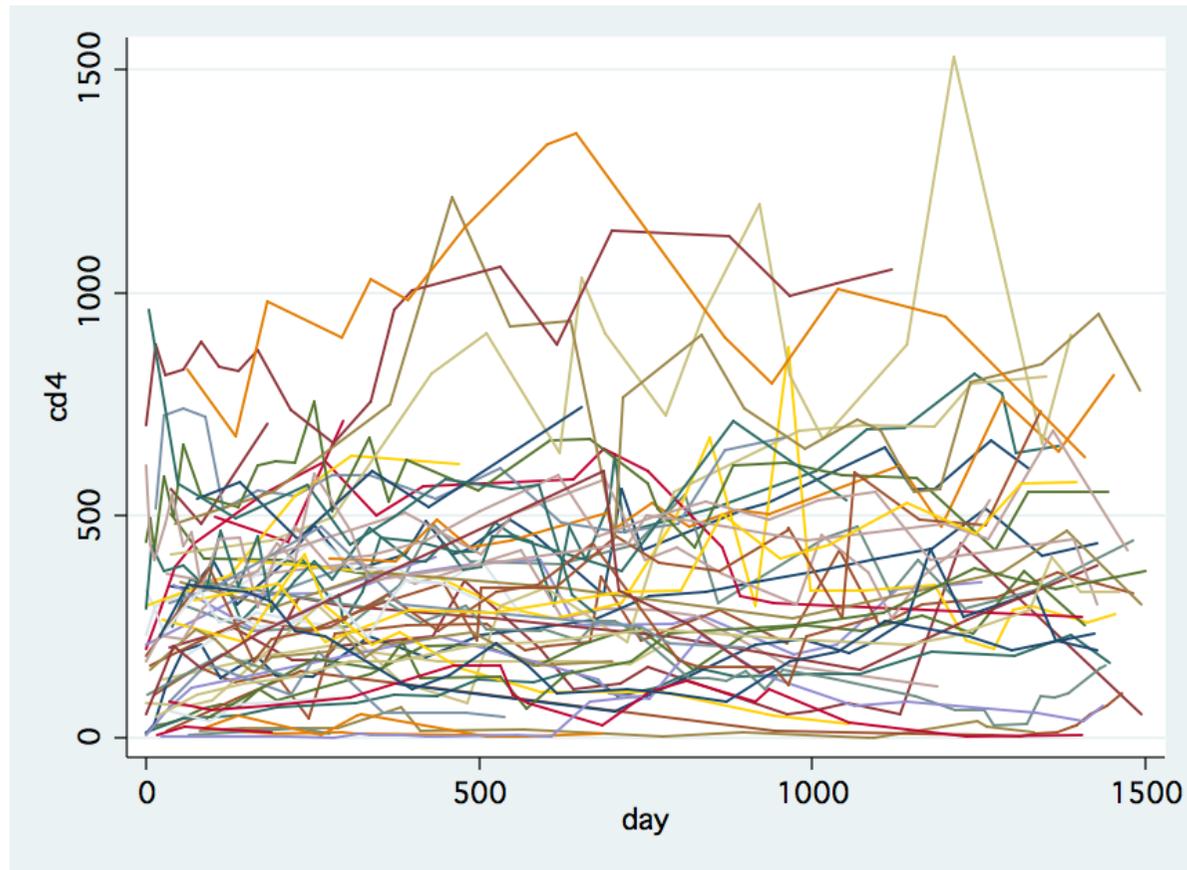
Strategia 1

Profileplot su campione casuale

- Procedura
 - Seleziona una percentuale di individui in maniera casuale
 - Traccia i profileplot degli individui corrispondenti a tale percentuale

Profileplot

Campione casuale = $0.10 * N$



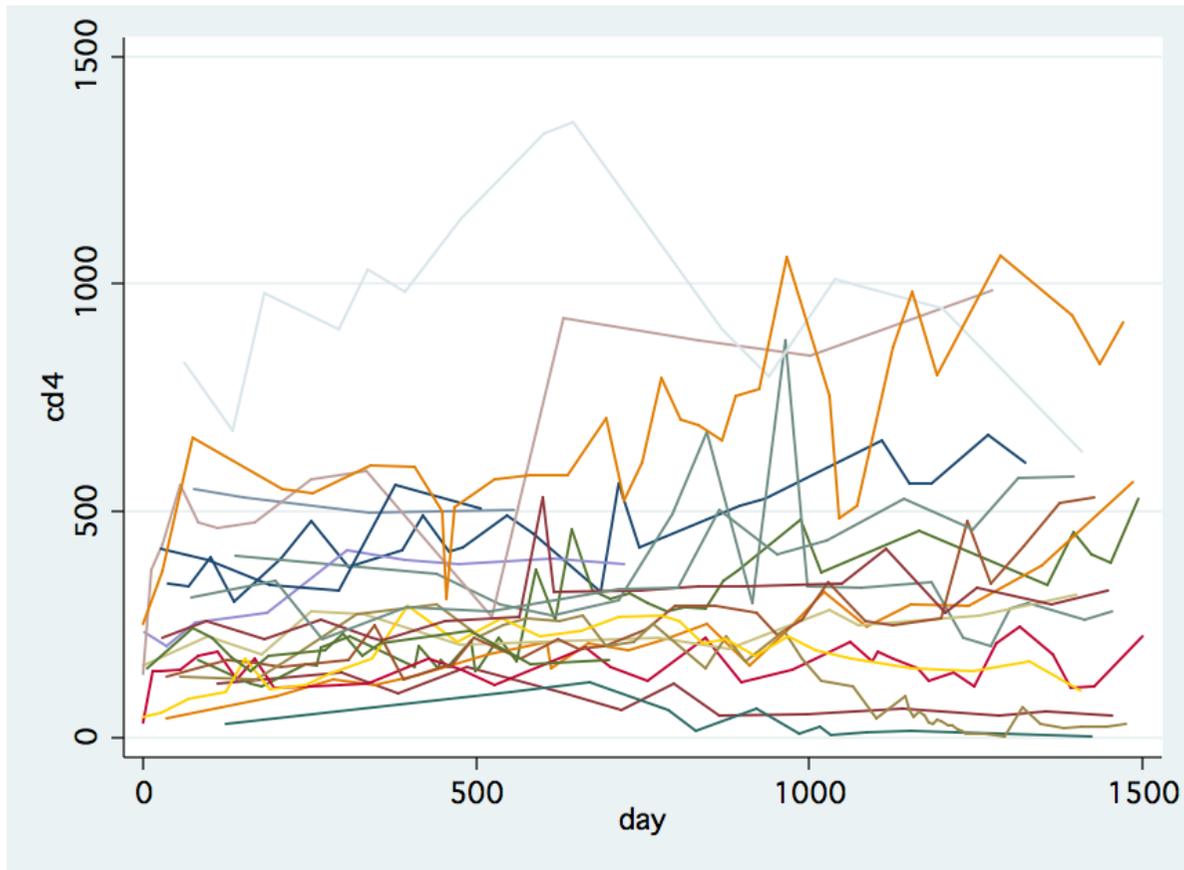
Strategia 2

Profileplot con ranking della media

- Procedura
 - Calcola la media dei CD4 per ogni soggetto
 - Traccia il profileplot di un numero predeterminato di soggetti scelti a caso e *ordinati* in base alla media
 - (Ovviamente si può fare lo stesso per la mediana ed eventualmente per misure di dispersione come la deviazione standard o il range interquartile)

Profileplot

Ranking della media, $N = 20$



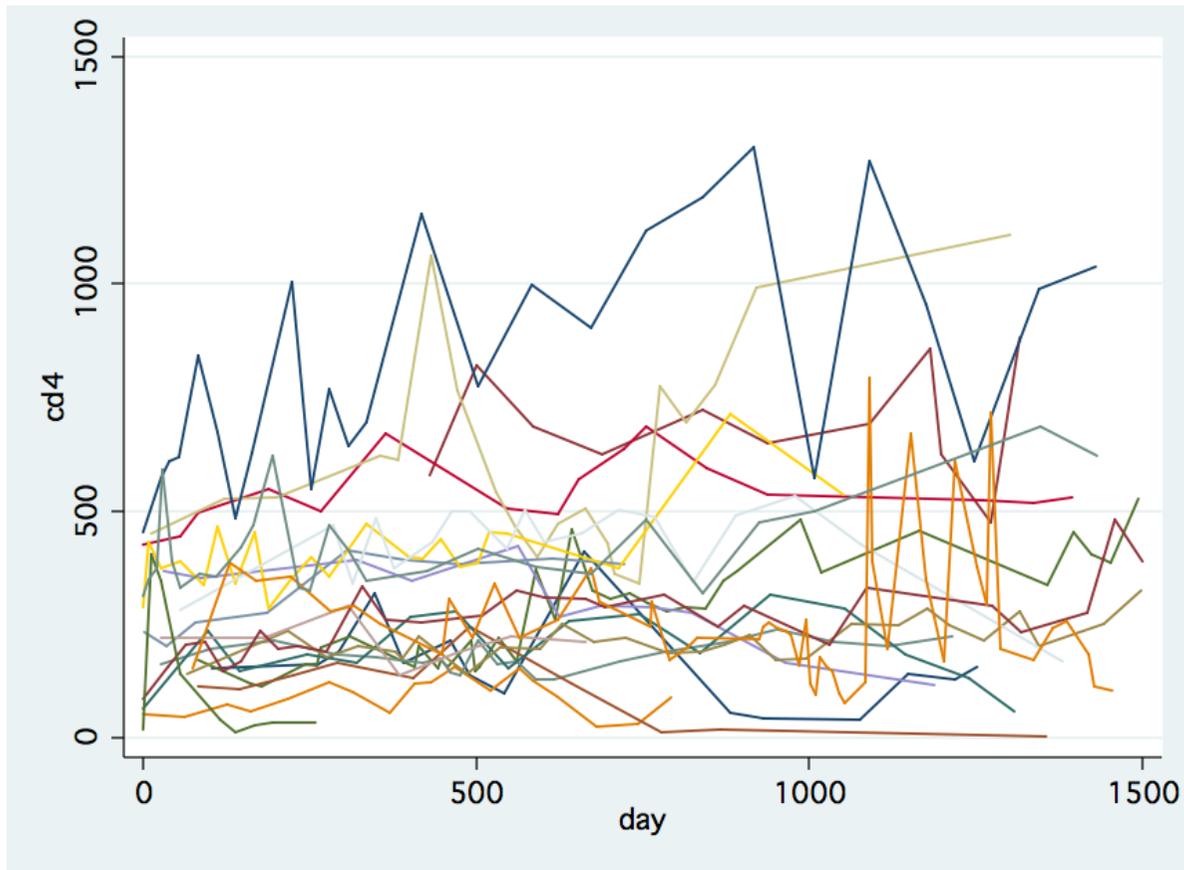
Strategia 3

Profileplot con ranking di AUC

- Procedura
 - Calcola l'AUC dei CD4 per ogni soggetto
 - Traccia il profileplot di un numero predeterminato soggetti scelti a caso e *ordinati* in base all'AUC

Profileplot

Ranking di AUC, $N = 20$

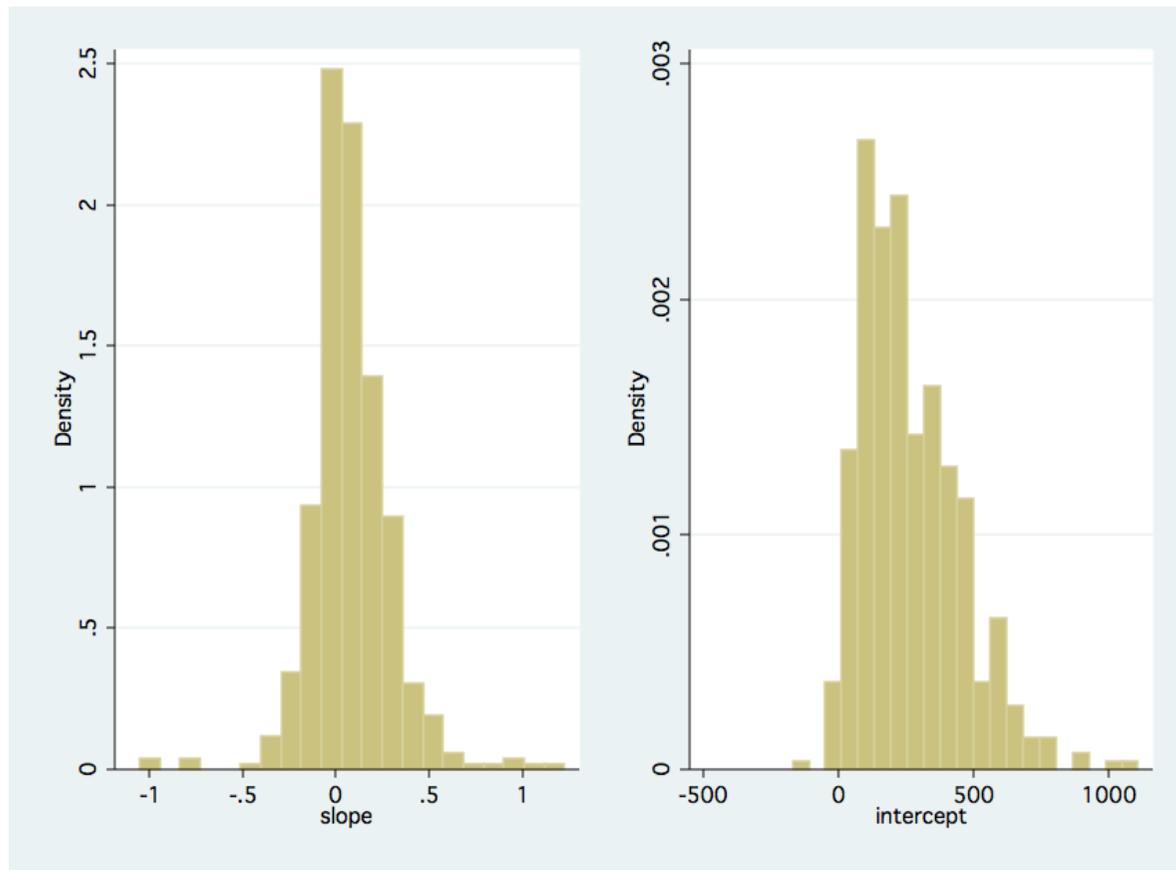


Strategia 4

Profileplot con ranking dello slope

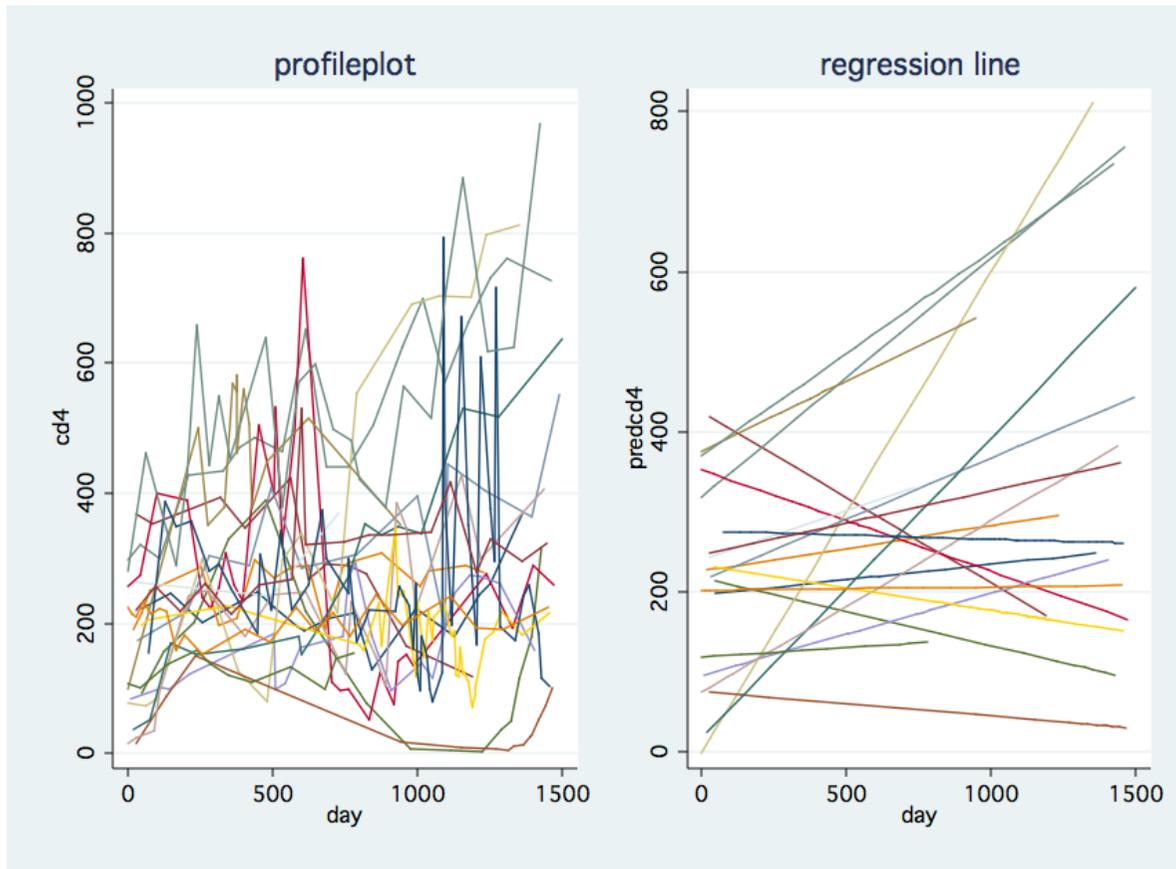
- Procedura
 - Calcola lo slope di CD4 vs. day per ogni individuo assumendo una relazione lineare (da testare!)
 - (Studia la distribuzione dello slope)
 - Traccia il profileplot di un numero predeterminato soggetti scelti a caso e *ordinati* in base allo slope

Distribuzione dello slope (e dell'intercetta)

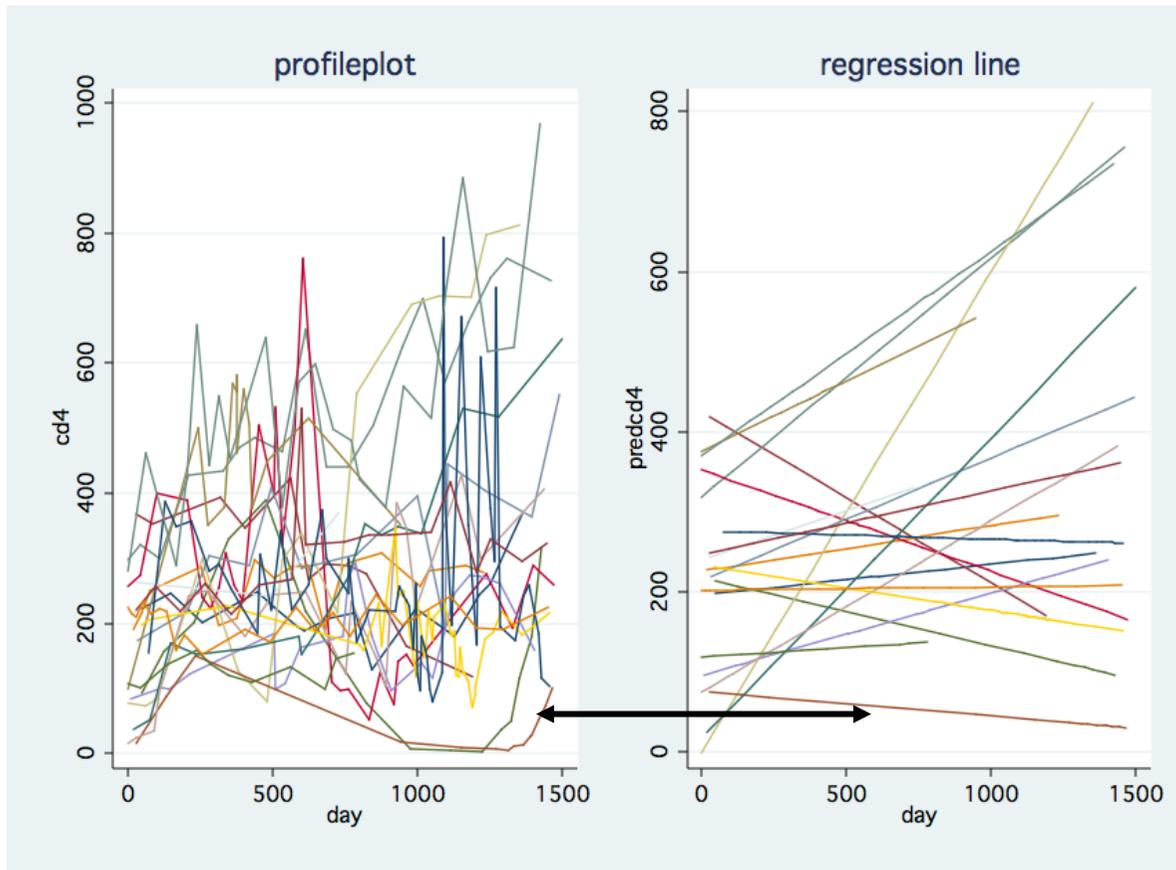


Profileplot

Ranking dello slope + retta predetta, $N = 20$



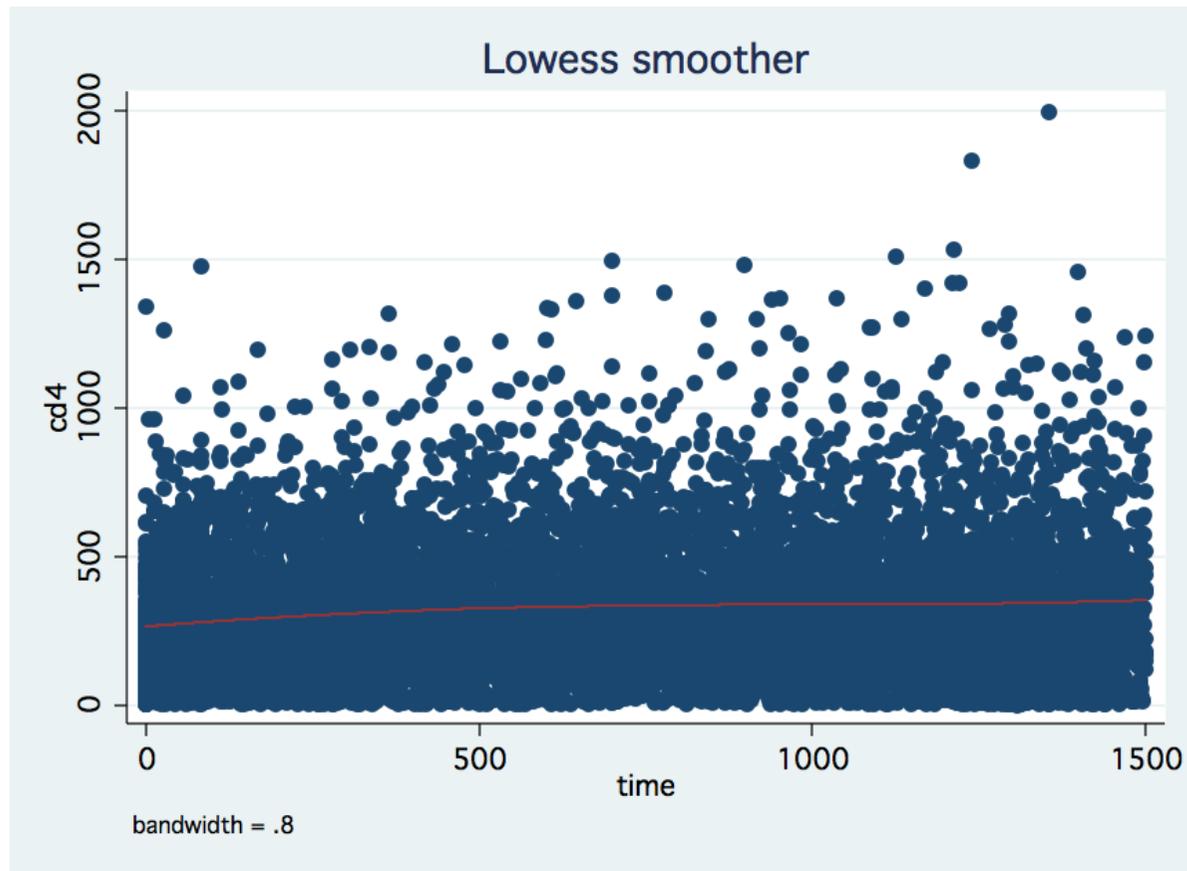
Confrontate queste due curve: appartengono allo stesso individuo



Metodi grafici non lineari

- Analizzando gli slope abbiamo assunto una relazione lineare tra Y (cd4) e t (year)
- In generale, conviene iniziare l'analisi utilizzando tecniche di *smoothing* che non assumono una relazione lineare tra Y e t

Locally weighted scatterplot smoothing (lowess)



Limiti

- Lo smoothing è stato effettuato su tutti i soggetti ma a noi interessa lo sviluppo individuale della relazione cd4-year
- Si tratta (ovviamente) dello stesso problema incontrato coi metodi grafici lineari e la soluzione è la stessa

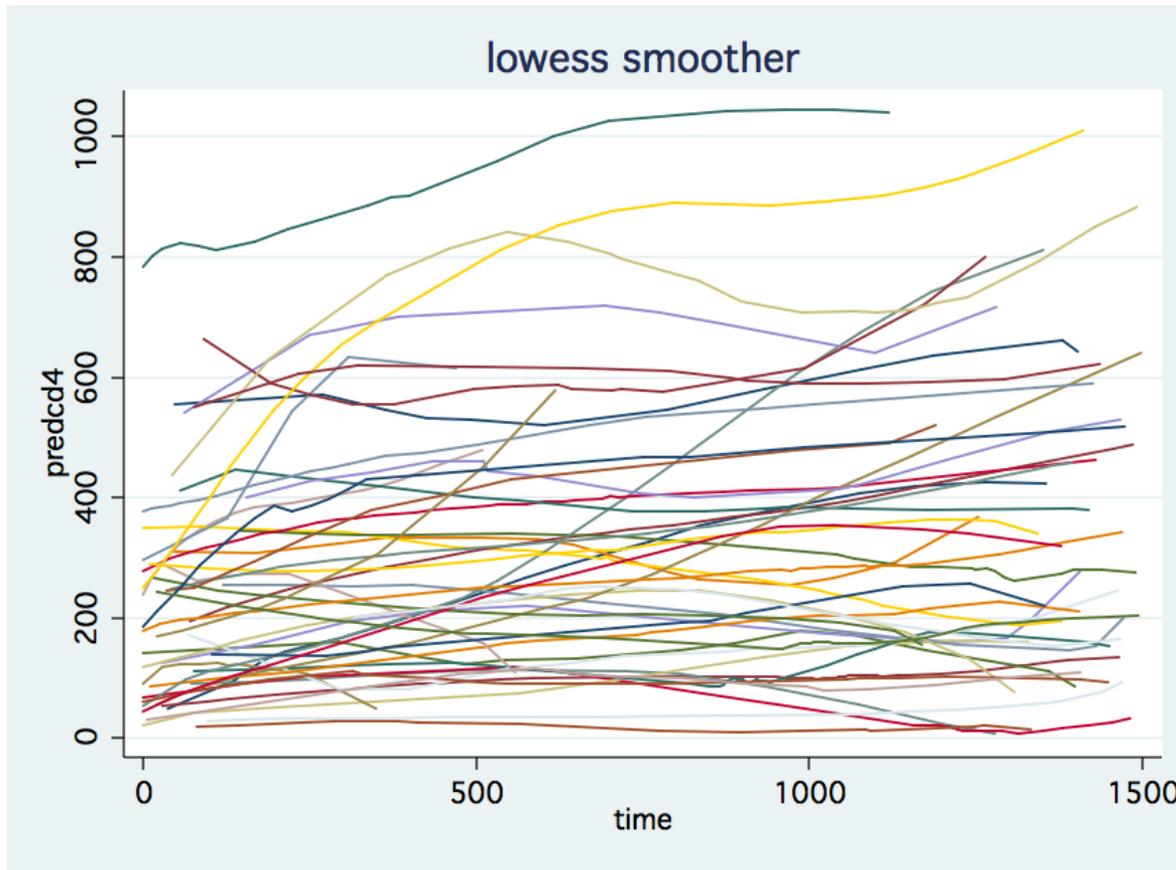
Strategia 5

Profileplot con lowess individuale

- Procedura
 - Stima i valori di CD4 sul tempo utilizzando il lowess smoother nei singoli individui
 - Disegna il profileplot di un numero predeterminato di individui scelti a caso

Strategia 5

Profileplot con lowess individuale, $0.10*N$



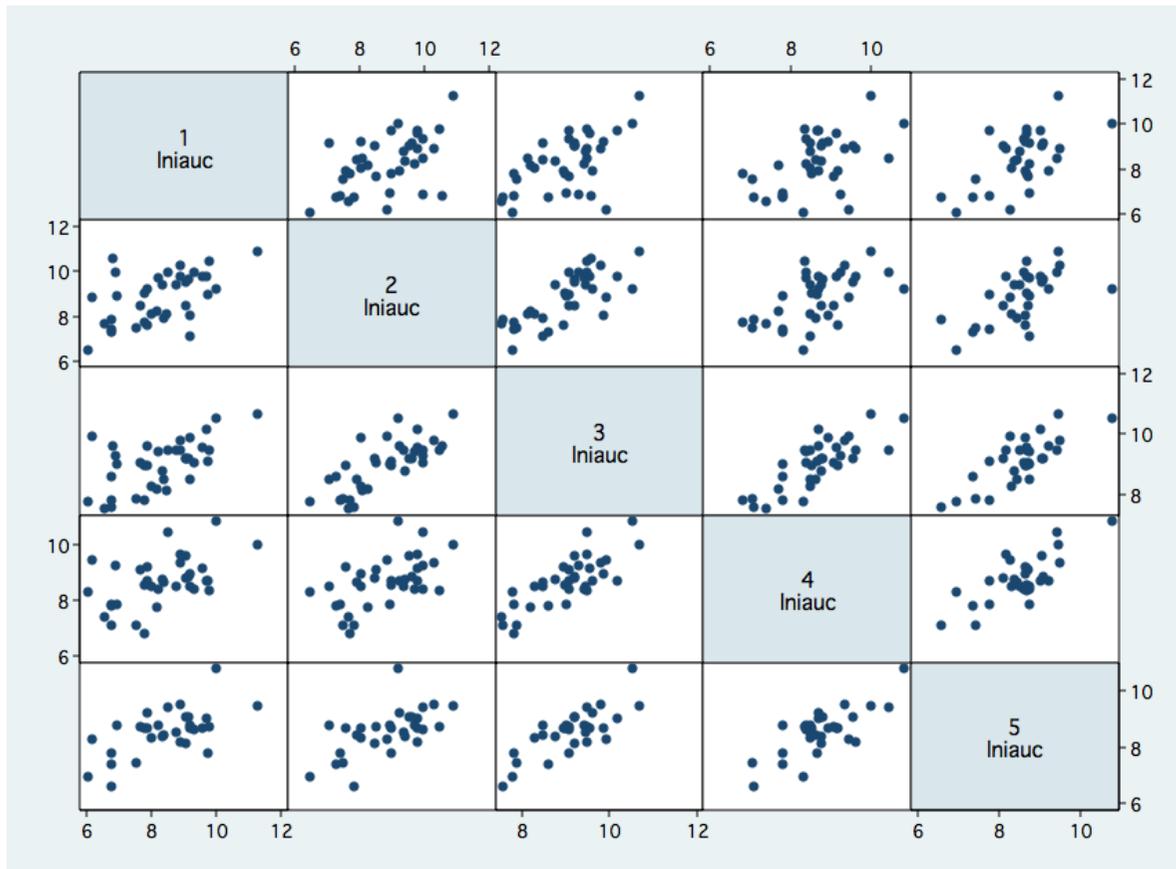
Altre strategie

- Ovviamente, possiamo fare un ordinamento per ranking dei lowess plots sulla base di altre variabili (come abbiamo fatto in precedenza)

Correlazione tra misure successive

- Come abbiamo detto, è altamente probabile che i valori dell'outcome (nonché dei predittori) dello stesso individuo siano correlati nel tempo
- Valutare preliminarmente il grado di correlazione è utile anche per la scelta della tecnica statistica da utilizzare per l'analisi

Matrice grafica ("pairwise scatterplot")



Matrice di correlazione "pairwise"

	lniauc1	lniauc2	lniauc3	lniauc4	lniauc5
lniauc1	1.0000 37				
lniauc2	0.5128 0.0012 37	1.0000 37			
lniauc3	0.5854 0.0001 37	0.7580 0.0000 37	1.0000 37		
lniauc4	0.4968 0.0024 35	0.5623 0.0004 35	0.7932 0.0000 35	1.0000 35	
lniauc5	0.6125 0.0002 31	0.6204 0.0002 31	0.7885 0.0000 31	0.7587 0.0000 31	1.0000 31

Sono riportati:

r di Pearson

valore di p

numero di soggetti (variabile)

(Inutile ricordarvi che stiamo continuando ad assumere una relazione lineare tra Y e t !)

Matrice di correlazione "listwise"

	lniauc1	lniauc2	lniauc3	lniauc4	lniauc5
lniauc1	1.0000 31				
lniauc2	0.6503 0.0001 31	1.0000 31			
lniauc3	0.6508 0.0001 31	0.7425 0.0000 31	1.0000 31		
lniauc4	0.5132 0.0032 31	0.5052 0.0037 31	0.7441 0.0000 31	1.0000 31	
lniauc5	0.6125 0.0002 31	0.6204 0.0002 31	0.7885 0.0000 31	0.7587 0.0000 31	1.0000 31

Sono riportati:
 r di Pearson
 valore di p
 numero di soggetti (variabile)

ANALISI DEI DATI LONGITUDINALI: SINOSI

Sinossi 1

	Metodi "tradizionali"		Metodi "moderni"	
<i>Outcome</i>	2 misurazioni	>2 misurazioni	2 misurazioni	>2 misurazioni
Continuo (gaussiana)	Test di Student Test di Wilcoxon	RM-MANOVA RM-ANOVA (regressione normale multivariata) Test di Friedman	Equazioni di stima generalizzata Modelli con effetti random Analisi del cambiamento (ANCOVA)	Equazioni di stima generalizzata Modelli con effetti random Altri modelli
Dicotomico (binomiale)	Test di McNemar "Proportion of change"	Test Q di Cochran "Proportion of change"	Modelli con effetti random Equazioni di stima generalizzata	Modelli con effetti random Equazioni di stima generalizzata
Conta (Poisson)	--	--	Equazioni di stima generalizzata Modelli con effetti random	Equazioni di stima generalizzata Modelli con effetti random

Sinossi 2

- Problemi particolari posti dagli studi longitudinali
 - Missing data
 - Le tecniche per “missing data” utilizzate nella letteratura biomedica sono spesso errate
 - Sample size
 - Le formule sono state sviluppate per i trial randomizzati controllati

ANALISI DI OUTCOME CONTINUI 1

Outcome continui

- Metodi tradizionali
- Metodi moderni

Metodi tradizionali

- Due misurazioni
 - Test t di Student per dati appaiati
(Student's t -test for paired data)
 - Test del segno dei ranghi per dati appaiati
(Wilcoxon matched-pairs signed-ranks test)
 - Cosa è invece il "sign test"?

Due misurazioni

- Ipotesi nulla per entrambi i test:
 - $H_0: (Y_{t2} - Y_{t1}) = 0$
- Ciò che ci interessa è la dimensione della differenza: il t -test può stimarla ma non il test di Wilcoxon
- Siete convinti che la dimensione e la variabilità dell'effetto sono più importanti del valore di p ?

Test t per dati appaiati

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{N}}}$$

d = media delle differenze

S_d = deviazione standard della
differenze

N = numero individui

Distribuzione t con $(n-1)$ gradi
di libertà

Con quale nome è indicato il
denominatore della statistica t ?

Cosa assume il t -test?

- Indipendenza dei valori di Y : cosa vuole dire?
- Distribuzione gaussiana della differenza d

Case-study

Summary for variables: iauc
by categories of: year

year	N	mean	sd	min	max
1	37	7915	13140	426	78379
2	37	11793	11971	646	53169
3	37	11036	9154	1866	42958
4	35	8523	9653	901	50040
5	31	7199	8363	734	48609
Total	177	9373	10720	426	78379

Case-study

L'iauc si modifica tra il primo e il secondo controllo?

(Non è una domanda particolarmente interessante ma dobbiamo cominciare dal piccolo per arrivare al grande...)

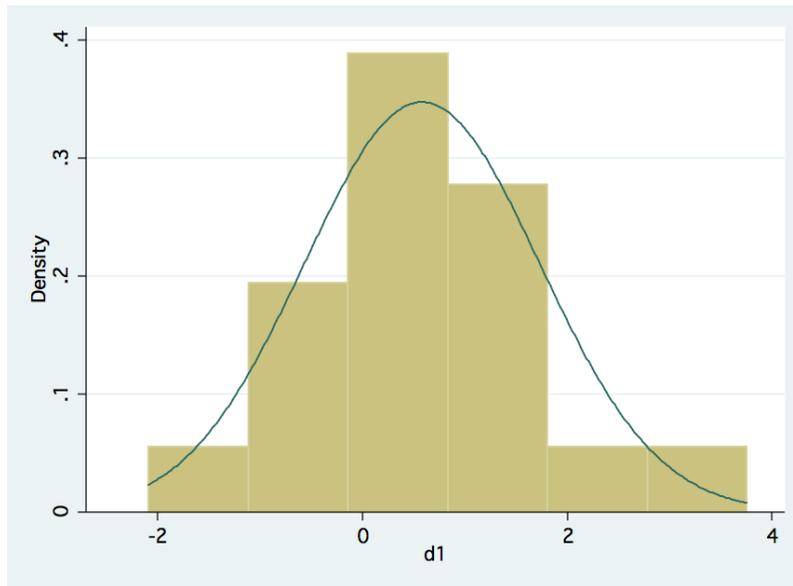
Test t per dati appaiati

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
iauc2	37	11793.16	1968.063	11971.26	7801.746	15784.58
iauc1	37	7915.386	2160.192	13139.93	3534.313	12296.46
diff	37	3877.776	1750.315	10646.75	327.9727	7427.58

$\text{mean}(\text{diff}) = \text{mean}(\text{iauc2} - \text{iauc1})$ $t = 2.2155$
 $H_0: \text{mean}(\text{diff}) = 0$ degrees of freedom = 36
 $H_a: \text{mean}(\text{diff}) < 0$ $H_a: \text{mean}(\text{diff}) \neq 0$ $H_a: \text{mean}(\text{diff}) > 0$
 $\text{Pr}(T < t) = 0.9834$ $\text{Pr}(|T| > |t|) = 0.0331$ $\text{Pr}(T > t) = 0.0166$

Ci ricordiamo di verificare le assunzioni?



Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
d1	37	0.96811	1.188	0.360	0.35941

Domande

- Siete stupiti del fatto che la differenza ($Y_{t2}-Y_{t1}$) ha una distribuzione (ragionevolmente) gaussiana mentre Y_1 e Y_2 non hanno affatto una distribuzione gaussiana?
- Cosa fate ora?

Caveat

- Il t -test è ragionevolmente robusto alle violazioni della normalità ma l'interpretazione della "differenza media" è complicata perché la media *non* è un buon indicatore del punto centrale ("mid-point") della distribuzione
- L'efficienza asintotica del test di Wilcoxon rispetto al t -test è 95.5% (95% nei "piccoli campioni")

Test di Wilcoxon per dati appaiati

- Valuta *sia* la direzione *sia* la grandezza della differenza
- Utile quando le assunzioni del t-test sono violate o non sono testabili (capita assai più spesso di quanto si creda)
- Praticamente indispensabile nei “piccoli campioni”, dove il *t*-test è inaffidabile
- Esiste una variante “esatta” (permutazione) del test di Wilcoxon (così come di molti altri test nonparametrici)

Problema

- Quanto deve essere numeroso un campione per essere considerato piccolo?
- (Benvenuti a una delle controversie più grandi della storia della statistica)
- E già che ci siamo: siamo sicuri che il “Teorema del limite centrale” autorizzi - come credono in molti - a trattare una distribuzione vistosamente non gaussiana come gaussiana?

Test di Wilcoxon per dati appaiati

- Procedura
 - Ordina le differenze positive in ranghi e sommale (T^+)
 - Ordina le differenze negative in ranghi e sommale (T^-)
 - (Controlla per le code) e sfrutta il fatto che la somma di tutti i ranghi è $N(N+1)/2$ sicché $T^- = N(N+1)/2 - T^+$

Test di Wilcoxon per dati appaiati

Wilcoxon signed-rank test

sign	obs	sum ranks	expected
positive	26	514	351.5
negative	11	189	351.5
zero	0	0	0
all	37	703	703

unadjusted variance **4393.75**
adjustment for ties **0.00**
adjustment for zeros **0.00**

adjusted variance **4393.75**

Ho: iauc2 = iauc1

z = **2.452**
Prob > |z| = **0.0142**

Da dove viene la "deviata standardizzata"??*

$$\mu_{T^+} = \frac{N(N+1)}{4}$$
$$\sigma_{T^+}^2 = \frac{N(N+1)(2N+1)}{24}$$
$$z = \frac{T^+ - \mu_{T^+}}{\sqrt{\sigma_{T^+}^2}}$$

T^+ ha una distribuzione
approssimativamente gaussiana
per $N > 15$

T^+ = somma dei ranghi positivi

N = numero individui

μ = media

σ^2 = varianza

*cacofonica traduzione italiana
di "z-score"

Metodi tradizionali

- Più di due misurazioni
 - Analisi multivariata della varianza per misure ripetute (RM-MANOVA)
 - Analisi univariata della varianza per misure ripetute (RM-ANOVA)
 - Test di Friedman per misure ripetute

RM-MANOVA

- La statistica è calcolata sulle $(T - 1)$ differenze tra le misure successive che vengono utilizzate simultaneamente come outcome
- (La MANOVA è una tecnica più generale che può essere utilizzata con outcome multivariati non necessariamente legati al tempo)

$(T-1) = (5-1) = 4$ differenze per lniauc

id	lniauc1	lniauc2	d1	lniauc3	d2	lniauc4	d3	lniauc5	d4
1	7.802388	9.005143	1.202755	8.966621	-.0385218	8.518114	-.4485073	8.689643	.1715288
2	6.758565	7.843163	1.084598	7.558292	-.2848711	7.083506	-.4747863	6.598937	-.4845691
3	8.452679	8.104797	-.3478813	8.131613	.0268154

d1 = (lniauc2-lniauc1)

d2 = (lniauc3-lniauc2)

d3 = (lniauc4-lniauc3)

d4 = (lniauc5-lniauc4)

*RM-MANOVA

$$F = \frac{N - T + 1}{(N - 1)(T - 1)} H^2$$

$$H^2 = \frac{N y'_d y_d}{S_d^2}$$

$$df = (T - 1), N(T + 1)$$

F = statistica F

N = numero soggetti

T = numero di misure ripetute

H² = test di Hotelling

y'_d = riga vettore delle differenze tra misure successive

df = gradi di libertà della statistica F

y_d = colonna vettore delle differenze tra misure successive

S²_d = matrice varianza/covarianza delle differenze tra misure successive

Intermezzo

- Definiamo insieme (anche senza formule!):
 - Varianza
 - Covarianza
 - Correlazione
 - Matrice di covarianza
 - Matrice di correlazione

Assunzioni della RM-MANOVA

- Le osservazioni ripetute dei differenti soggetti sono indipendenti
- Le osservazioni hanno una distribuzione normale multivariata

Limiti della RM-MANOVA

- Non fornisce nessuna metrica *biologicamente* o *clinicamente* utile della dimensione dell'effetto
- L'andamento dell'outcome nel tempo può essere inferito solo visivamente
- Può modellare effetti non-lineari del tempo soltanto utilizzando contrasti predeterminati
- Non consente di modellare covariate tempo-dipendenti

Limiti della RM-MANOVA

- Non può modellare covariate continue (comune a tutte le forme di ANOVA)
- Non può gestire dati "unequally spaced", a meno di non ricorrere a discutibili definizioni di finestre temporali (comune a tutte le forme di ANOVA)
- Non gestisce "internamente" i missing data

Conclusione sulla RM-MANOVA

- L'RM-MANOVA ha attualmente un interesse esclusivamente storico e questa è la ragione per cui ho scelto di non riportare alcun esempio
- È davvero *impensabile* utilizzarla per studi clinici dove l'obiettivo principale è la stima della dimensione dell'effetto in relazione al tempo e alle altre covariate di interesse

RM-ANOVA

- C'è un solo outcome, non più i $(T-1)$ outcome della RM-MANOVA
- Il tempo è ovviamente modellato come variabile categorica

Assunzioni della RM-ANOVA

- Le osservazioni ripetute dei differenti soggetti sono indipendenti
- Le osservazioni hanno una distribuzione normale multivariata
- *Le varianze delle differenze degli stessi individui sono le stesse (sfericità o "sphericity")*
- (La sfericità è sempre presente se tutte le varianze e le covarianze sono uguali, ovvero se c'è "compound symmetry")

Assunzioni della RM-ANOVA

- L'assenza di sfericità può essere controllata utilizzando appositi "fattori di correzione" e ciò viene fatto in automatico dalla quasi totalità dei software
- Si tratta di testare l'ipotesi nulla che la sfericità (ϵ) è uguale a 0
- Se l'ipotesi è falsificata, viene introdotta una correzione per la sfericità

Case Study

id	year				
	1	2	3	4	5
1	7.802388	9.005143	8.966621	8.518114	8.689643
2	6.758565	7.843163	7.558292	7.083506	6.598937
3	8.452679	8.104797	8.131613		
4	9.199144	8.068555	9.877455	8.947229	8.670809
5	9.707885	9.794126	10.1634	8.695541	9.041416
6	8.176023	8.251916	8.186704	7.714271	
7	6.798174	10.5503	9.589787		
8	9.732687	8.965747	9.085687	8.66708	7.785794
9	6.054508	6.471098	7.763533	8.305324	6.949126
10	7.544367	7.494416	7.864146	7.069607	7.427398
11	11.26931	10.88123	10.66799	9.979335	9.484181
12	9.058519	9.554308	9.196788	9.565461	9.07126
13	6.922096	8.931167	9.023047	7.821874	8.770679
14	9.995975	9.214225	10.53673	10.82058	10.79157
15	9.157388	9.629995	9.182408	8.813157	9.08931
16	6.871688	9.97414	9.280853	9.240447	
17	9.178376	7.090023	8.486117	8.482439	8.760248
18	8.762021	9.384687	9.449035	8.488068	8.510432
19	7.884237	9.222177	9.601895	8.695361	9.23868
20	6.169966	8.842792	9.922088	9.437631	8.298246
21	6.764805	7.283782	8.594929	7.802472	7.364487
22	8.495903	9.961084	9.476467	10.42902	9.421758
23	9.759712	10.47589	9.483161	8.351232	8.697862
24	9.036681	8.483968	9.193432	8.765714	8.101644
25	8.88103	9.799477	9.474699	9.610456	8.183962
26	8.206142	9.703087	9.429483	8.38602	8.776166
27	7.883738	7.598686	8.934992	9.188642	8.666317
28	6.777114	7.422827	7.818245	7.825264	7.760626
29	6.536905	7.69066	7.531791	7.383675	
30	8.353855	9.399694	8.763591	8.759775	8.374022
31	7.792697	7.705675	7.806584	6.803592	
32	8.012866	8.072335	8.277271	8.503084	8.335947
33	9.561052	9.803889	9.558519	9.123484	8.694735
34	8.38666	7.918072	8.485402	8.627706	8.447584
35	8.900544	10.29953	9.795848	9.346352	9.490349
36	7.647504	8.501854	9.062358	9.066505	8.734669
37	9.310816	9.954126	9.05653	8.381988	8.640035

Case Study

- Il valore di $\ln iauc$ si modifica con il passare del tempo?
- Ipotesi nulla corrispondente all'analisi della varianza (ANOVA):

$$H_0 = \overline{\ln iauc_1} = \overline{\ln iauc_2} = \overline{\ln iauc_3} = \overline{\ln iauc_4} = \overline{\ln iauc_5}$$

Domande

- Come esprimete questa ipotesi nulla utilizzando le differenze?
- Cosa significa che l'ANOVA è un "omnibus test"?
- Come faccio a stabilire le differenze tra un punto e l'altro o tra un certo numero di punti di interesse? (In parte abbiamo già risposto a questa domanda parlando della RM-MANOVA)

Perché questa ANOVA non è corretta?

Number of obs = 177 R-squared = 0.0665
Root MSE = .987456 Adj R-squared = 0.0448

Source	Partial SS	df	MS	F	Prob > F
Model	11.9521009	4	2.98802523	3.06	0.0180
year	11.9521009	4	2.98802523	3.06	0.0180
Residual	167.711937	172	.975069399		
Total	179.664038	176	1.0208184		

Perché questa ANOVA è incompleta?

Number of obs = 177 R-squared = 0.7002
Root MSE = .629322 Adj R-squared = 0.6120

Source	Partial SS	df	MS	F	Prob > F
Model	125.801745	40	3.14504362	7.94	0.0000
id	113.849644	36	3.16249011	7.99	0.0000
year	12.9636659	4	3.24091647	8.18	0.0000
Residual	53.8622928	136	.39604627		
Total	179.664038	176	1.0208184		

Ecco la RM-ANOVA: cosa mancava?

Number of obs = 177 R-squared = 0.7002
 Root MSE = .629322 Adj R-squared = 0.6120

Source	Partial SS	df	MS	F	Prob > F
Model	125.801745	40	3.14504362	7.94	0.0000
id	113.849644	36	3.16249011	7.99	0.0000
year	12.9636659	4	3.24091647	8.18	0.0000
Residual	53.8622928	136	.39604627		
Total	179.664038	176	1.0208184		

Between-subjects error term: id
 Levels: 37 (36 df)
 Lowest b.s.e. variable: id

Repeated variable: year

Huynh-Feldt epsilon = 0.7526
 Greenhouse-Geisser epsilon = 0.6897
 Box's conservative epsilon = 0.2500

Source	df	F	Prob > F			
			Regular	H-F	G-G	Box
year	4	8.18	0.0000	0.0001	0.0001	0.0072
Residual	136					

Limiti della RM-ANOVA

- Quali sono i limiti comuni la RM-MANOVA?
- In quali casi la RM-ANOVA è preferibile alla RM-MANOVA?
- Adesso passate pure alla prossima diapositiva!

Should we quit using repeated measures analysis of variance?

Repeated Measures ANOVA, R.I.P.?



Charles E. McCulloch

Chance 2005;18:29.

Requiescat in pacem?

“If the data are from a simple design and very close to balanced, it is nice to be able to use simpler techniques that have exact distributions under exact balance and exact normality.

In such a case, repeated measures ANOVA is the approach of choice.”

McCulloch CE *Chance* 2005;18:29.

Requiescat in pacem?

“However, my opinion is that, in the vast majority of cases, the newer mixed model algorithms are by far the better choices.

These methods allow full and efficient use of the data, do not require manual specification of test statistics, allow very flexible correlation structures, allow inferences about the correlation and covariance structure, and generate better predicted values than ANOVA-based methods.”

McCulloch CE *Chance* 2005;18:29.

Idea: e se usassimo la regressione normale multivariata?

- Quali vantaggi avremmo rispetto alla ANOVA multivariata?
- Ha il vantaggio teorico di parametri calcolabili da formule precise ("closed form") e con distribuzioni statistiche note
- Ha il limite pratico di complicarsi notevolmente (= stima della matrice di varianza-covarianza) all'aumentare dei predittori
- In conclusione: non è una grande idea...

Test di Friedman

$$H = \frac{12 \sum_{t=1}^T R_t^2}{NT(T+1)} - 3N(T+1)$$

H = statistica di Friedman

R = somma dei ranghi

t = tempo

N = numero di soggetti

T = numero di misure ripetute

Test di Friedman per misure ripetute

- Può testare solo l'ipotesi che Y non si modifica in t (disegno "one-within")
- Non fornisce indicazioni biologicamente o clinicamente utili sull'andamento di Y nel tempo
- Come confrontiamo due o più punti tra loro?

MODULO 2

Modulo 2

- Analisi di outcome continui 2
- Equazioni generalizzate di stima
- Modelli con effetti random
- Altri modelli
- Analisi del cambiamento (2 punti)

ANALISI DI OUTCOME CONTINUI 2

Metodi

- Più di due misurazioni
 - Altri metodi tradizionali oltre a quelli considerati nel Modulo 1
 - Moderni

Altri metodi tradizionali

- Riducono il “problema longitudinale” a uno o più problemi trasversali
- Indispensabili fino a quando non vennero teorizzati e utilizzati i metodi moderni (anni 80)
- Considereremo criticamente i principali metodi

Cosa non funziona con questi metodi "tradizionali" 1 ?

1. Uso l'ultimo punto disponibile dell'outcome e una statistica riassuntiva del predittore come la media
2. Faccio un'analisi del cambiamento utilizzando solo due punti (es. inizio e fine). Nota bene: se ho *solo* due punti, questo metodo è ottimo (lo analizzeremo in dettaglio più avanti)

Cosa non funziona con questi metodi "tradizionali" 2 ?

3. Utilizzo lo slope dell'outcome sul tempo come variabile dipendente e lo slope del predittore sul tempo come variabile indipendente
4. Utilizzo lo slope dell'outcome sul tempo come variabile dipendente e il valore del predittore al basale come variabile indipendente

Metodi moderni

- Equazioni generalizzate di stima (GEE, generalized estimating equations)
- Modelli con effetti random
- Altri modelli

Un modello generale di regressione

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \varepsilon_{it}$$

Y = outcome

i = individuo

t = tempo

β_0 = intercetta

j = numero di predittori

β_{1j} = slope del predittore j

X_{itj} = predittore

ε_{it} = errore

Coefficiente di regressione

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \varepsilon_{it}$$

β_{1j} misura la "relazione longitudinale" tra l'outcome Y e il predittore X_j

Modellare l'effetto del tempo è facile!

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t + \varepsilon_{it}$$

Con questa notazione il tempo t è una covariata come gli altri predittori

EQUAZIONI GENERALIZZATE DI STIMA

Le origini

Biometrika (1986), **73**, 1, pp. 13–22
Printed in Great Britain

Longitudinal data analysis using generalized linear models

BY KUNG-YEE LIANG AND SCOTT L. ZEGER

*Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205,
U.S.A.*

SUMMARY

This paper proposes an extension of generalized linear models to the analysis of longitudinal data. We introduce a class of estimating equations that give consistent estimates of the regression parameters and of their variance under mild assumptions about the time dependence. The estimating equations are derived without specifying the joint distribution of a subject's observations yet they reduce to the score equations for multivariate Gaussian outcomes. Asymptotic theory is presented for the general class of estimators. Specific cases in which we assume independence, m -dependence and exchangeable correlation structures from each subject are discussed. Efficiency of the proposed estimators in two simple situations is considered. The approach is closely related to quasi-likelihood.

Cosa fanno le GEE?

- Assumono una struttura di correlazione intratemporale tra l'outcome e il predittore
- Effettuano la stima utilizzando una procedura "quasi-likelihood"

Cosa è il quasi-likelihood?

- Il metodo principale di stima utilizzato in statistica è il “maximum likelihood” (MLE, maximum likelihood estimation)
- Questo metodo richiede che si specifichi una distribuzione statistica per i dati
- Il quasi-likelihood richiede che si specifichi (soltanto) una struttura per la media e una relazione tra la media e la varianza
- In molti casi è efficiente come il likelihood e richiede meno assunzioni (ciò non significa che sia una panacea!)

Le GEE secondo il “modello generale”

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t + \mathbf{CORR}_{it} + \varepsilon_{it}$$

- \mathbf{CORR}_{it} = struttura o matrice di correlazione (“working correlation structure”, “working correlation matrix”)

“Independent structure”

	t_1	t_2	t_3	t_4	t_5
t_1	—	0	0	0	0
t_2	0	—	0	0	0
t_3	0	0	—	0	0
t_4	0	0	0	—	0
t_5	0	0	0	0	—

Non c'è associazione tra le misure successive (!)

A cosa può servire una struttura del genere (oltre che a fini didattici?)

“Exchangeable structure” 1

	t_1	t_2	t_3	t_4	t_5
t_1	—	ρ	ρ	ρ	ρ
t_2	ρ	—	ρ	ρ	ρ
t_3	ρ	ρ	—	ρ	ρ
t_4	ρ	ρ	ρ	—	ρ
t_5	ρ	ρ	ρ	ρ	—

C'è la stessa correlazione (ρ)
tra le misure successive

Qual è il vantaggio di questa
struttura?

Abbiamo già incontrato
qualcosa di simile?

“Exchangeable structure” 2

	t_1	t_2	t_3	t_4	t_5
t_1	—	ρ	ρ	ρ	ρ
t_2	ρ	—	ρ	ρ	ρ
t_3	ρ	ρ	—	ρ	ρ
t_4	ρ	ρ	ρ	—	ρ
t_5	ρ	ρ	ρ	ρ	—

Se (e solo se) l'outcome è continuo, una GEE con questa struttura corrisponde a un modello con intercetto random (lo considereremo tra breve)

“(Stationary) m-dependent structure”

	t_1	t_2	t_3	t_4	t_5
t_1	—	ρ_1	ρ_2	0	0
t_2	ρ_1	—	ρ_1	ρ_2	0
t_3	ρ_2	ρ_1	—	ρ_1	ρ_2
t_4	0	ρ_2	ρ_1	—	ρ_1
t_5	0	0	ρ_2	ρ_1	—

Le correlazioni distanti t sono uguali, così come quelle distanti $t+1$, fino a $t+m$

È riportata una matrice di correlazione con $m = 2$, che pertanto si chiamerà... ?

“Autoregressive structure”

	t_1	t_2	t_3	t_4	t_5
t_1	—	ρ^1	ρ^2	ρ^3	ρ^4
t_2	ρ^1	—	ρ^1	ρ^2	ρ^3
t_3	ρ^2	ρ^1	—	ρ^1	ρ^2
t_4	ρ^3	ρ^2	ρ^1	—	ρ^1
t_5	ρ^4	ρ^3	ρ^2	ρ^1	—

Le correlazioni distanti $t=1$
sono uguali, così come quelle
distanti $t=2$, fino a $t=T$

“Unstructured correlation”

	t_1	t_2	t_3	t_4	t_5
t_1	–	ρ^1	ρ^2	ρ^3	ρ^4
t_2	ρ^1	–	ρ^6	ρ^7	ρ^8
t_3	ρ^2	ρ^6	–	ρ^{10}	ρ^{11}
t_4	ρ^3	ρ^7	ρ^{10}	–	ρ^{13}
t_5	ρ^4	ρ^8	ρ^{11}	ρ^{13}	–

I coefficienti di correlazione sono differenti

È la matrice di correlazione meno restrittiva, ma ha un limite vistoso: quale?

Altre strutture di correlazione

- Noi abbiamo considerato solo le matrici di correlazione più utilizzate
- (Considereremo altre possibili strutture di varianza-covarianza più avanti nel modulo alla voce "ALTRI MODELLI")

Come scelgo la matrice di correlazione?

- Idealmente, andrebbe specificata prima dell'analisi
- È sempre utile "visualizzare" la matrice di correlazione intra-individuale *post-hoc*
- La matrice migliore è quella che impiega il minor numero di parametri
- Sono disponibili dei test per identificare la miglior matrice (si basano un criterio di informazione di Akaike modificato)

Come scelgo la matrice di correlazione?

- Le GEE sono considerate “robuste” alla misspecificazione della matrice di correlazione ma ciò non è sempre vero
- Potete cautelarvi utilizzando intervalli di confidenza “robusti” (il cosiddetto “sandwich estimator” di Huber)
- In verità, questi intervalli sono “semi-robusti” perché *non* sono robusti a un errore di specificazione della struttura della media

Un'obiezione metodologica

“This working correlation matrix is not necessarily assumed to actually described the correlation matrix of the data.

Instead, a fix is applied to the estimated correlation matrix of the fixed effects to correct the standard errors.

This is known as the ‘sandwich estimator’”

Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005, p. 398.

*Un'obiezione metodologica

“I do not like statistical models that do not completely specify the sampling distribution of the data, or worse, assume that you have made a mistake in specifying the distribution of the data and then try to fix it.

My preference is, if the model for the data is not correct, then I attempt to correct it and develop a model as accurate as possible”

Weiss RE. *Modeling longitudinal data*. New York: Springer, 2005, p. 398.

Contro-obiezioni?

- Io non sono d'accordo
- Prima di dirvi perché la penso diversamente, ditemi cosa ne pensate voi

Torniamo a Iniauc

Estimated within-id correlation matrix R:

	c1	c2	c3	c4	c5
r1	1				
r2	.5803724	1			
r3	.4135987	.7399763	1		
r4	.5064082	.5319523	.5613584	1	
r5	.5916477	.512319	.4320354	.4740032	1

- Che differenza c'è tra questa matrice e quella che abbiamo visto discutendo della visualizzazione dei dati longitudinali? Suggerimento: leggete la prima riga
- Quale struttura è implicita in questa matrice?
- Pensate che la matrice cambierà inserendo altri predittori oltre al tempo?

Come interpreto β_1 ?

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} + \dots$$

β_1 combina la variabilità intra-individuale e quella inter-individuale

Il modello che ne risulta è detto *marginale* proprio per questo motivo

Pertanto, non è il metodo da utilizzare se interessa valutare la variabilità intraindividuale

Una vecchia polemica

STATISTICS IN MEDICINE, VOL. 17, 447–469 (1998)
Statist. Med., 17, 447–469 (1998)

ON THE APPROPRIATENESS OF MARGINAL MODELS FOR REPEATED MEASUREMENTS IN CLINICAL TRIALS

J. K. LINDSEY^{1*} AND P. LAMBERT²

¹*Biostatistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium*

²*Université de Liège, Faculté d'Economie, de Gestion et de Sciences Sociales, Liège, Belgium*

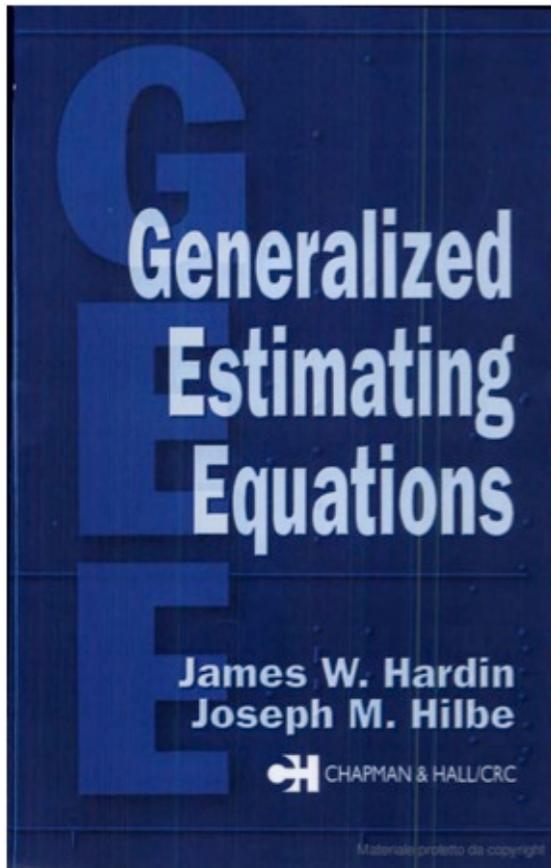
SUMMARY

Although models developed directly to describe marginal distributions have become widespread in the analysis of repeated measurements, some of their disadvantages are not well enough known. These include producing profile curves that correspond to no possible individual, possibly showing that a treatment is superior on average when it is poorer for each individual subject, implicitly generating complex and implausible physiological explanations, including underdispersion in subgroups, and sometimes corresponding to no possible probabilistic data generating mechanism. We conclude that such marginal models may sometimes be appropriate for descriptive observational studies, such as sample surveys in epidemiology, but should only be used with great care in causal experimental settings, such as clinical trials. © 1998 John Wiley & Sons, Ltd.

Nota tecnica

- Le GEE sono un'estensione dei modelli lineari generalizzati (GLM)
- Cosa è un GLM? Quali sono i vantaggi di "programmare" un modello come GLM?

Testo consigliato



Hardin JW *et al.* Generalized estimating equations. Boca Raton: Chapman & Hall/CRC; 2003.

Ottima trattazione tecnica delle GEE.

Case Study

Long-Term Evaluation of Glucose Homeostasis in a Cohort of HAART-Treated HIV-Infected Children

A Longitudinal, Observational Cohort Study

Alessandra Viganò,¹ Paolo Brambilla,² Giulia Pattarino,¹ Sara Stucchi,¹ Silvia Fasan,¹ Chiara Raimondi,¹ Chiara Cerini,¹ Vania Giacomet,¹ Gian V. Zuccotti¹ and Giorgio Bedogni³

1 Pediatric Clinic, 'L. Sacco' Hospital, University of Milan, Milan, Italy

2 Azienda Sanitaria Locale Milano 2, Milan, Italy

3 Clinical Epidemiology Unit, Liver Research Centre, Basovizza, Trieste, Italy

Case Study

Background and objectives: Few and mainly cross-sectional studies of glucose homeostasis are available in HIV-infected children treated with highly active antiretroviral therapy (HAART). The aim of the present study was to describe a 4-year course of glucose homeostasis in a cohort of HAART-treated children and adolescents, using glucose and insulin levels during an oral glucose tolerance test (OGTT) as outcome measures. In addition, we investigated possible risk factors, both related and unrelated to antiretroviral therapy, associated with insulin resistance.

Case Study

Table II. Predictors of the changes in the natural logarithm (\log_e)-transformed area under the serum concentration-time curve (AUC) of insulin during oral glucose tolerance testing

Variable	\log_e AUC insulin regression coefficient (95% CI)
Ritonavir/lamivudine/stavudine vs indinavir/lamivudine/stavudine	-0.29 (-0.85, 0.26)
Nelfinavir/lamivudine/stavudine vs indinavir/lamivudine/stavudine	-0.69** (-1.29, -0.09)
Efavirenz/lamivudine/tenofovir vs indinavir/lamivudine/stavudine	-0.93** (-1.82, -0.03)
Time (y)	0.58*** (0.21, 0.94)
Time ² (y ²)	-0.13*** (-0.20, -0.05)
Drug*time interaction	0.07 (-0.02, 0.15)
Male sex	-0.18 (-0.54, 0.18)
Age at baseline (y)	0.01 (-0.03, 0.04)
Tanner stage ≥ 4	1.32† (0.91, 1.74)
z-BMI (SD score) ^a	0.13 (-0.07, 0.32)
CD4+ T cells (%)	-0.01 (-0.02, 0.01)
Constant	7.50† (6.50, 8.50)

a Compared with US reference data.^[25]

AUC = area under the serum concentration-time curve; **BMI** = body mass index; **SD** = standard deviation; **z** = difference between the individual value and the population mean, divided by the population SD. ** $p < 0.05$; *** $p < 0.01$; † $p < 0.001$.

*Case Study

```

GEE population-averaged model
Group and time vars:      id time
Link:                     identity
Family:                   Gaussian
Correlation:              unstructured
Scale parameter:         .5480879
Number of obs            =      177
Number of groups        =       37
Obs per group: min     =        3
                      avg     =   4.5
                      max     =    5
Wald chi2(11)          =   85.75
Prob > chi2            =   0.0000

```

(Std. Err. adjusted for clustering on id)

lniauc	Semi-robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
_Ipi_2	-.2945769	.2821982	-1.04	0.297	-.8476751	.2585214
_Ipi_3	-.6904682	.307598	-2.24	0.025	-1.293349	-.0875872
_Ipi_4	-.9250489	.4586728	-2.02	0.044	-1.824031	-.0260667
piXtime	.065243	.0444618	1.47	0.142	-.0219006	.1523867
time	.5775095	.1861518	3.10	0.002	.2126587	.9423602
time2	-.1296648	.0381169	-3.40	0.001	-.2043725	-.0549572
male	-.1801819	.1821742	-0.99	0.323	-.5372369	.176873
age1	.0068882	.0185007	0.37	0.710	-.0293726	.043149
tanner	1.323826	.2100434	6.30	0.000	.9121485	1.735504
zbmius	.1289238	.0995917	1.29	0.195	-.0662724	.3241199
cd4100	-.0063478	.0090987	-0.70	0.485	-.024181	.0114853
_cons	7.503327	.5098689	14.72	0.000	6.504002	8.502651

MODELLI CON EFFETTI RANDOM

Le origini

BIOMETRICS 38, 963–974
December 1982

Random-Effects Models for Longitudinal Data

Nan M. Laird^{1,2} and James H. Ware¹

¹Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A. and ²Department of Statistics, Harvard University, Science Center 603, 1 Oxford Street, Cambridge, Massachusetts 02138, U.S.A.

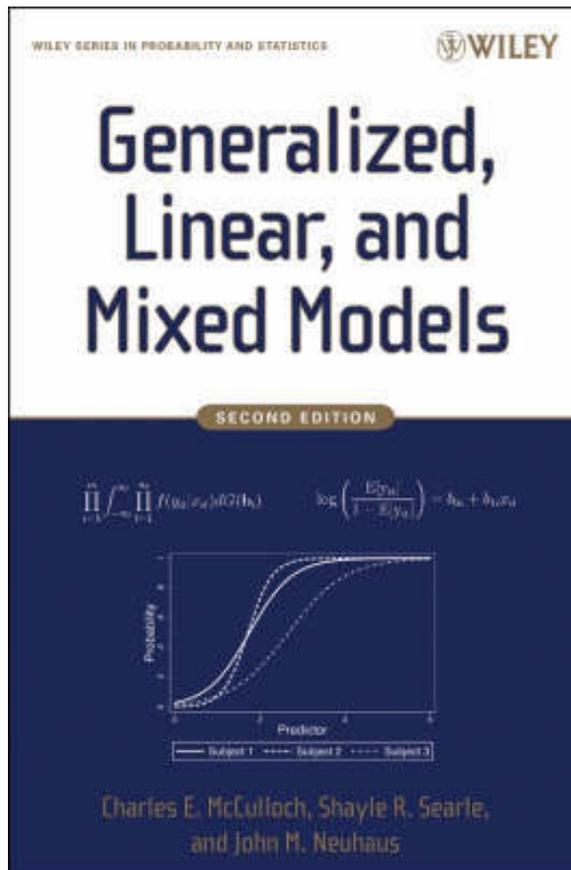
SUMMARY

Models for the analysis of longitudinal data must recognize the relationship between serial observations on the same unit. Multivariate models with general covariance structure are often difficult to apply to highly unbalanced data, whereas two-stage random-effects models can be used easily. In two-stage models, the probability distributions for the response vectors of different individuals belong to a single family, but some random-effects parameters vary across individuals, with a distribution specified at the second stage. A general family of models is discussed, which includes both growth models and repeated-measures models as special cases. A unified approach to fitting these models, based on a combination of empirical Bayes and maximum likelihood estimation of model parameters and using the EM algorithm, is discussed. Two examples are taken from a current epidemiological study of the health effects of air pollution.

Sinonimi

- Modelli con effetti random ("Random effect models")
 - Perché includono "effetti random"
- Modelli misti ("Mixed models")
 - Perché includono sia "effetti fissi" sia "effetti random"
- Modelli multivello ("Multilevel models")
 - Perché ci sono più livelli di analisi e l'analisi longitudinale è solo *una* delle applicazioni
- Modelli gerarchici ("Hierarchical models")
 - Sinonimo di modelli multivello

Testo consigliato

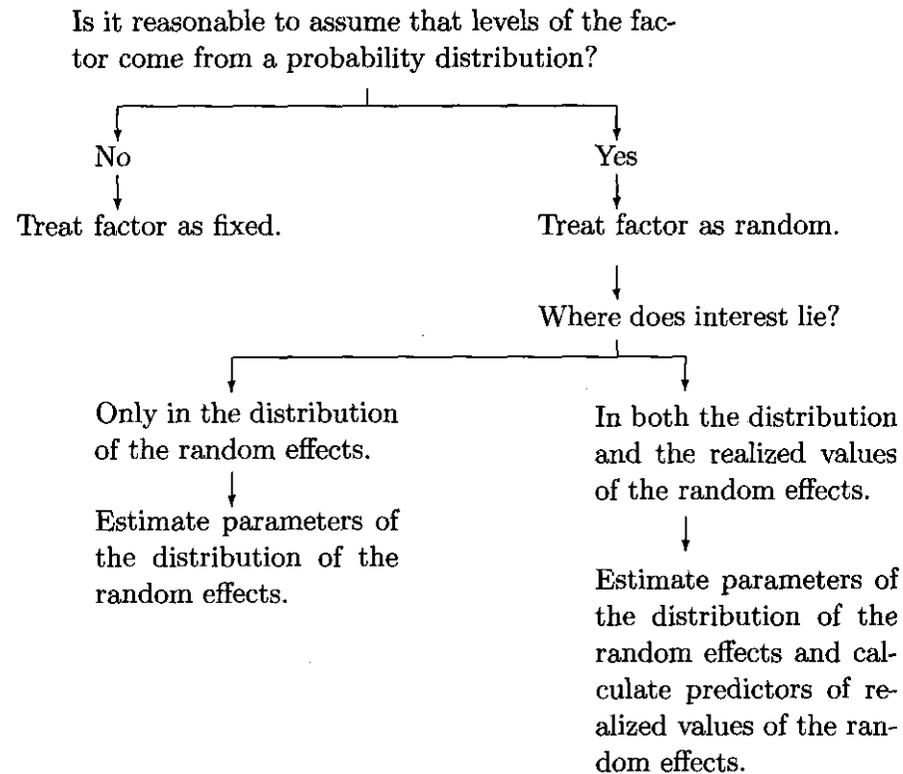


McCulloch CE *et al.*
*Generalized, linear, and mixed
models.* NY: Wiley; 2008.

Ottima introduzione tecnica ai
modelli misti (e molto di più!).

Cosa è un effetto random?

Figure 1.2: Decision tree for deciding fixed versus random



McCulloch CE *et al.* Generalized, linear, and mixed models. NY: Wiley; 2008, p. 19.

Modello con intercetto random
"Random intercept model"

$$Y_{it} = \beta_{0i} + \beta_1 t + \varepsilon_{it}$$

Ciascun individuo ha un "suo" intercetto β_{0i}

Modello con slope random "Random slope model"

$$Y_{it} = \beta_0 + \beta_{1i}t + \varepsilon_{it}$$

Ciascun individuo ha un "suo" slope β_{1i} associato al tempo

Nota: un modello con slope random senza un intercetto random *non* ha applicazioni utili e lo presentiamo unicamente per ragioni didattiche

*Modello con coefficienti random
"Random coefficients model"

$$Y_{it} = \beta_{0i} + \beta_{1i}t + \varepsilon_{it}$$

Ciascun individuo ha un "suo" slope β_{1i} e un "suo" intercetto β_{0i}

(NB: in un modello multilivello ci può essere più di un intercetto/
slope)

Metodi di stima per i modelli misti

- Il maximum likelihood (ML) è il metodo principale di stima utilizzato in statistica
- REML (restricted maximum likelihood) è un metodo che consente stime *indipendenti* dagli effetti fissi
- Ci sono diverse ragioni teoriche per preferire REML a ML ma, all'atto pratico, la differenza è poco rilevante
- Tutte le stime dei nostri case-studies sono ottenute con ML

Modelli con effetti random

- Per comprendere i vantaggi (e i limiti) dei modelli con effetti random, è necessario ricorrere a una notazione matematica (leggermente) più rigorosa

Modello con intercetto random

$$y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij}$$

i = individuo

j = tempo

u = intercetto random

ε = errore

Assunzioni:

ε ha una distribuzione
gaussiana con varianza τ^2

u ha una distribuzione
gaussiana con varianza σ^2

ε ed u non sono correlati tra i
differenti soggetti e alle
covariate x

Modello con intercetto random

$$\text{Var}(u_i + \varepsilon_{ij}) = \sigma^2 + \tau^2$$

La varianza *residua* totale è la somma della varianza residua intra-individuale σ^2 e della varianza residua inter-individuale τ^2

$$\text{Cov}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) = \tau^2$$

La covarianza tra i residui totali a due punti di tempo j e j' sullo stesso soggetto è τ^2 (indotta dall'intercetto random)

Modello a intercetto random

$$\text{Cor}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) = \frac{\tau^2}{\sigma^2 + \tau^2}$$

Il coefficiente di correlazione intraclass (ICC, "intraclass correlation coefficient" o ρ , rho) misura la proporzione della varianza residua totale spiegata dall'individuo

Perché questo modello è detto anche "variance components"?

Modello a coefficienti random

$$y_{ij} = x'_{ij}\beta + u_{0i} + u_{1i}t_{ij} + \varepsilon_{ij}$$

i = individuo

j = tempo

u_0 = intercetto random

u_1 = slope random

ε = errore

Assunzioni:

Gli effetti random hanno una distribuzione normale bivariata con varianze t_0^2 e t_1^2 e covarianza t_{01}

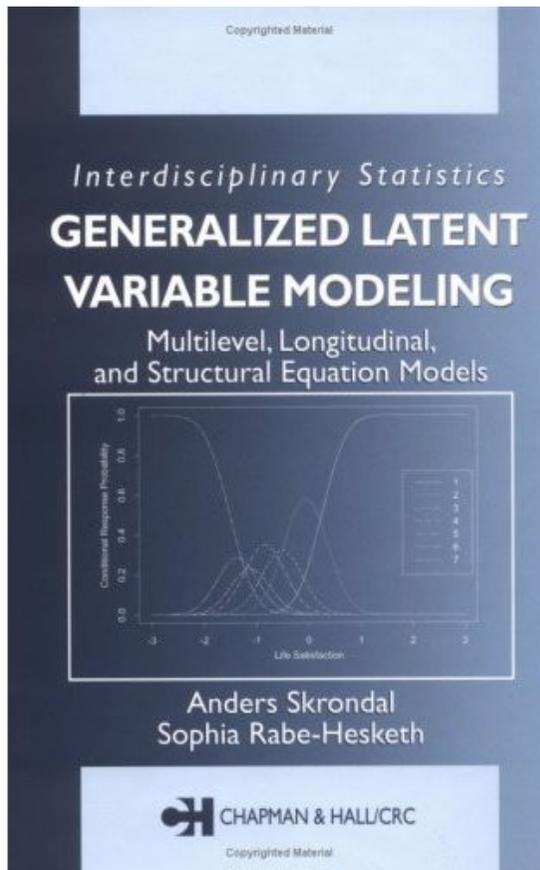
Gli effetti random non sono correlati tra gli individui, con l'errore ε e con le covariate x_{ij}

Modello a coefficienti random

$$\text{Var}(u_{0i} + u_{1i}t_{ij} + \varepsilon_{ij}) = \left(\tau_0^2 + 2\tau_{01}t_{ij} + \tau_1^2 t_{ij}^2 \right) + \sigma^2$$

- A differenza del modello a intercetto random, quello a coefficienti random non è più costante nel tempo
- Da notare che *sia* la varianza dell'intercetto random *sia* la correlazione tra i coefficienti random dipendono dalla t_{ij}

Testo consigliato



- Skrondal A *et al.* *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton: Chapman & Hall; 2004.

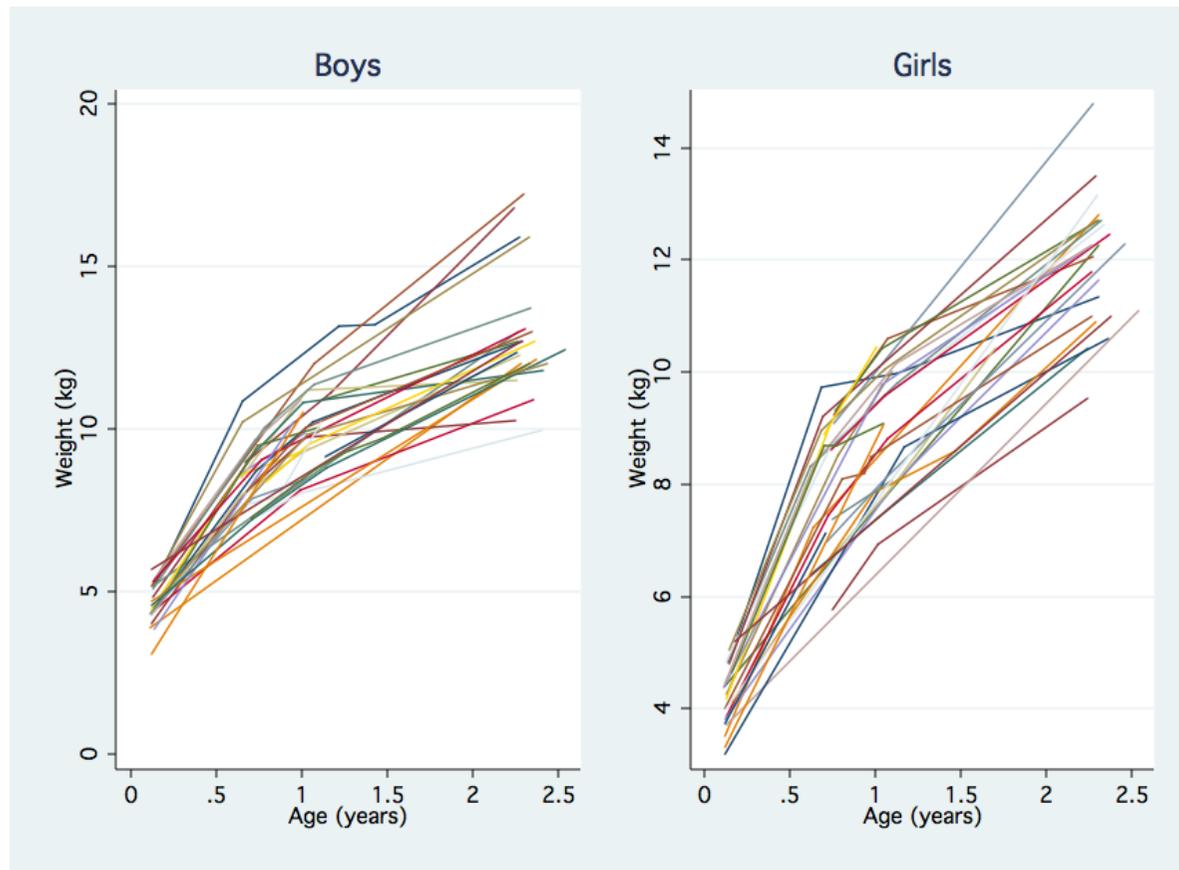
Ottima discussione tecnica dei modelli multilivello e longitudinali (e molto di più!) con splendidi esempi pratici

Case-study

- 60 bambini misurati in 5 occasioni (nascita, 6 settimane, 8 mesi, 12 mesi e 27 mesi)
- La variabile tempo è rappresentata dall'età
- Ci sono differenza nello "sviluppo" del peso all'aumentare dell'età? Il sesso ha un effetto su tale "sviluppo"?
- Ci sono missing data "gestiti" dalla procedura ML utilizzata per il calcolo (più dettagli in seguito)

Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modelling using STATA. Stata Press; 2008. www.stata-press.com/data/mlmus2.

Profileplots per sesso



Modello a intercetto random

```

Random-effects ML regression
Group variable: id

Random effects u_i ~ Gaussian

Log likelihood = -276.83266

Number of obs   =   198
Number of groups =    68

Obs per group: min =    1
                  avg =   2.9
                  max =    5

LR chi2(2)      =   445.86
Prob > chi2     =    0.0000
    
```

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.817918	.2900197	26.96	0.000	7.24949	8.386346
age2	-1.705599	.1087079	-15.69	0.000	-1.918663	-1.492536
_cons	3.432859	.1811825	18.95	0.000	3.077748	3.787971
/sigma_u	.9182256	.0973788			.7458966	1.130369
/sigma_e	.7347063	.0452564			.6511507	.8289837
rho	.6096746	.0614028			.4859915	.7230959

Likelihood-ratio test of sigma_u=0: `chibar2(01)= 78.07` Prob>=chibar2 = 0.000

$age2 = age^2$

Esercizio: calcoliamo il valore di rho (ICC) da questo output

Aggiungo un "random slope" per l'età: cosa cambia ?

```

Mixed-effects ML regression          Number of obs   =   198
Group variable: id                  Number of groups =    68

                                     Obs per group: min =    1
                                     avg =       2.9
                                     max =       5

Log likelihood = -258.07784          Wald chi2(2)    =  1978.20
                                     Prob > chi2      =   0.0000
    
```

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.703998	.2394082	32.18	0.000	7.234767	8.173229
age2	-1.660465	.0885229	-18.76	0.000	-1.833967	-1.486963
_cons	3.494512	.1372636	25.46	0.000	3.22548	3.763544

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Unstructured				
var(age)	.2540969	.0886513	.1282406	.5034694
var(_cons)	.4044398	.1645247	.1822166	.8976763
cov(age,_cons)	.0880874	.0880255	-.0844393	.2606142
var(Residual)	.331517	.0582668	.2349094	.467855

LR test vs. linear regression: chi2(3) = 115.58 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

Confronto formale dei due modelli

Likelihood-ratio test
(Assumption: `ri` nested in `.`)

LR $\chi^2(2)$ = 37.51
Prob > χ^2 = 0.0000

Note: LR test is conservative

- Il modello a coefficienti random è "nidificato" ("nested") dentro a quello a coefficienti random per cui posso utilizzare un likelihood-ratio test
- In generale, come posso confrontare modelli non nidificati?

Aggiungo il sesso come "effetto fisso": cosa succede?

```

Mixed-effects ML regression      Number of obs   =   198
Group variable: id              Number of groups =    68

                                Obs per group: min =    1
                                avg   =    2.9
                                max   =    5

                                Wald chi2(3)   =  1975.44
                                Prob > chi2    =   0.0000

Log likelihood = -253.86692

```

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.697967	.2382121	32.32	0.000	7.23108	8.164855
age2	-1.657843	.0880529	-18.83	0.000	-1.830423	-1.485262
male	.5960093	.1963689	3.04	0.002	.2111332	.9808853
_cons	3.19876	.1668333	19.17	0.000	2.871773	3.525747

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Unstructured				
var(age)	.2598032	.0888719	.1328834	.5079469
var(_cons)	.3537049	.1534276	.1511517	.8276928
cov(age,_cons)	.0476259	.0875144	-.1238993	.219151
var(Residual)	.3275618	.0568065	.2331718	.4601616

LR test vs. linear regression: chi2(3) = 104.17 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

C'è "interazione" tra sesso ed età?

```

Mixed-effects ML regression      Number of obs   =   198
Group variable: id              Number of groups =    68

                                Obs per group: min =    1
                                avg   =    2.9
                                max   =    5

                                Wald chi2(4)   =  2023.53
                                Prob > chi2    =  0.0000

Log likelihood = -252.99486
  
```

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.584403	.2530566	29.97	0.000	7.088421	8.080384
age2	-1.658569	.0879161	-18.87	0.000	-1.830881	-1.486257
male	.5040552	.20718	2.43	0.015	.0979898	.9101206
maleXage	.2303089	.1731562	1.33	0.183	-.109071	.5696889
_cons	3.244552	.169803	19.11	0.000	2.911744	3.57736

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Unstructured				
var(age)	.2469558	.0870363	.1237726	.4927358
var(_cons)	.3469542	.1521094	.1469244	.8193141
cov(age,_cons)	.0547442	.0856815	-.1131885	.222677
var(Residual)	.328304	.0571464	.2334056	.4617861

LR test vs. linear regression: chi2(3) = 104.77 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

Visualizzazione e predizione

- Una volta che abbiamo stimato gli slope e gli intercetti random possiamo utilizzarli per:
 - Visualizzare il modello e valutare le sue assunzioni
 - Predire valori utilizzando il modello

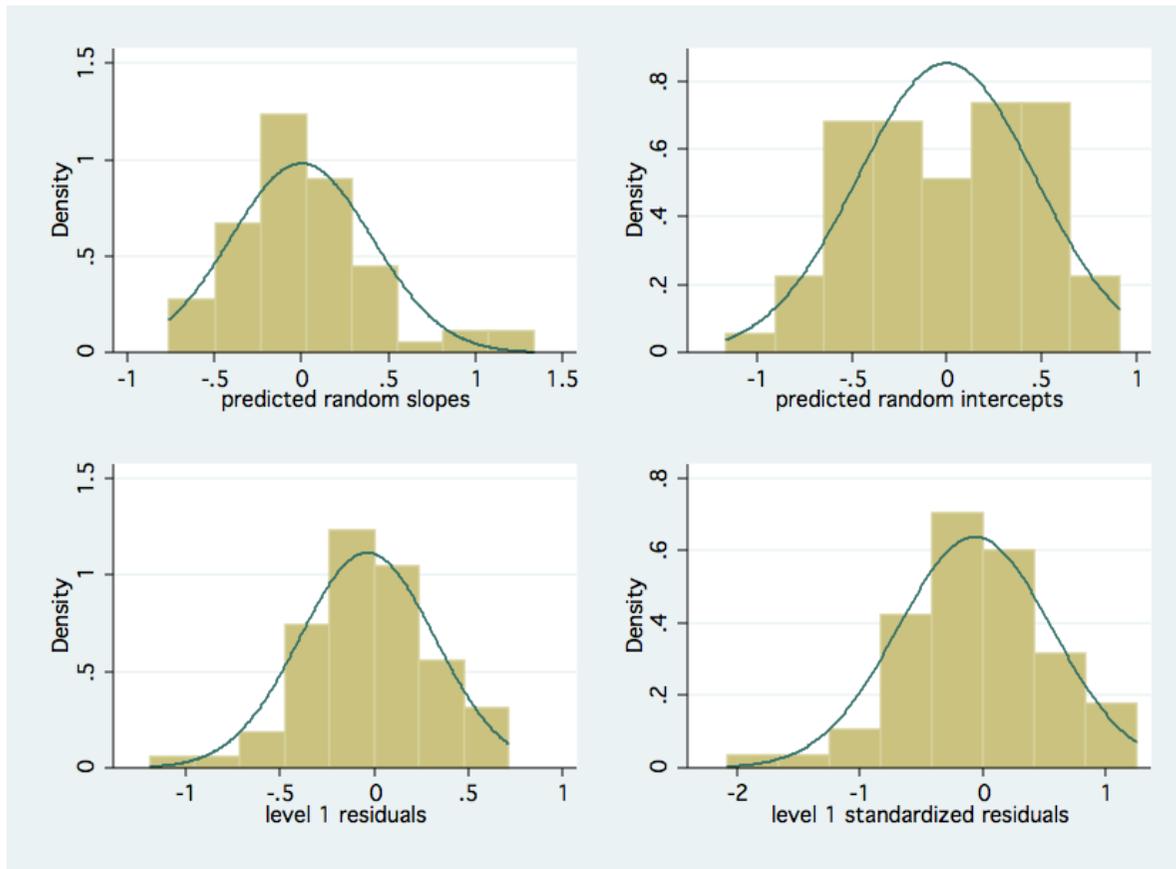
*Visualizzazione e predizione

- La procedura migliore per calcolare gli effetti random non è MLE ma l' "empirical Bayes prediction" (EBP)
- Notate l'aggettivo "empirical": non è un metodo bayesiano ma frequentista
- L'EBP riduce la varianza in ragione di uno "shrinkage" verso la media
- (Cosa è il bias-variance tradeoff e perché è importante quando si elaborano dei modelli predittivi?)

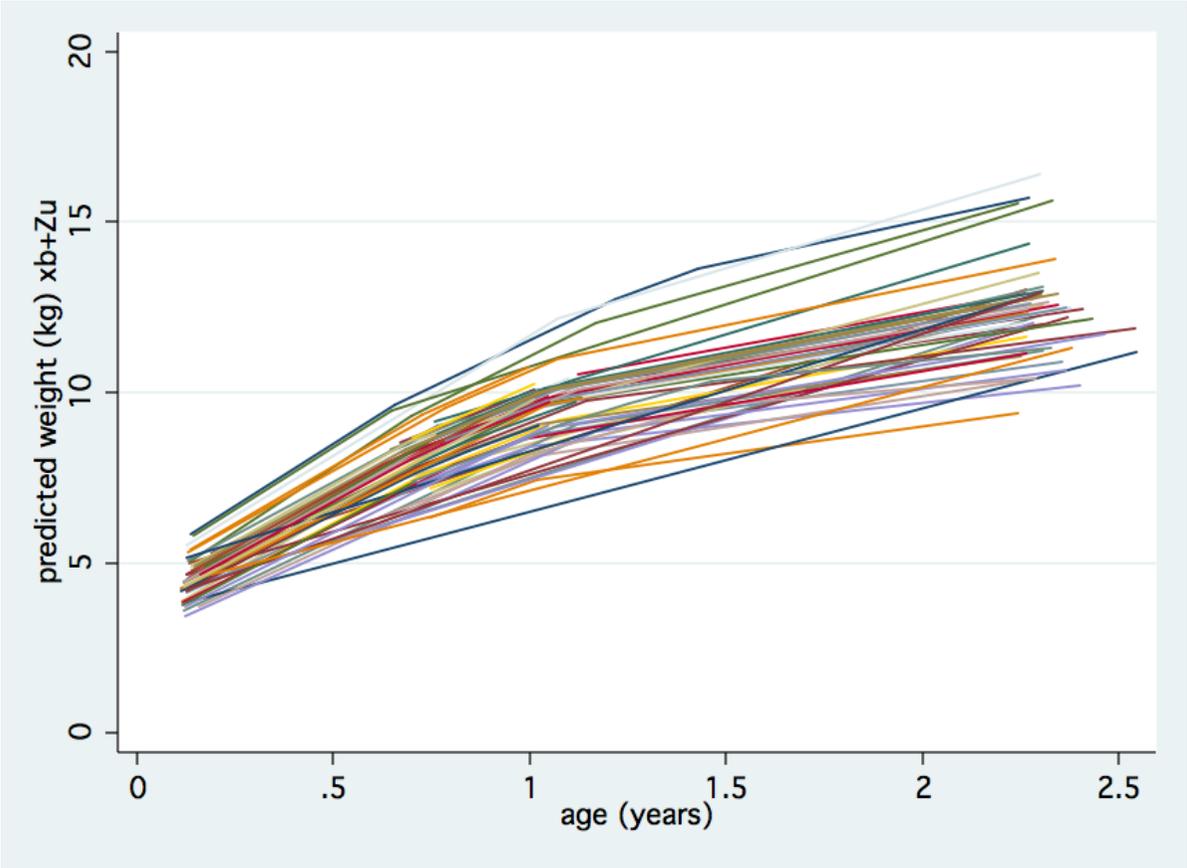
Diagnostica del modello

- Se le assunzioni di “normalità” per lo slope random, l’intercetto random e i residui di livello 1 sono soddisfatte, anche i valori predetti dall’EBP devono essere normalmente distribuiti
- Quindi, possiamo testare le assunzioni utilizzando questi parametri

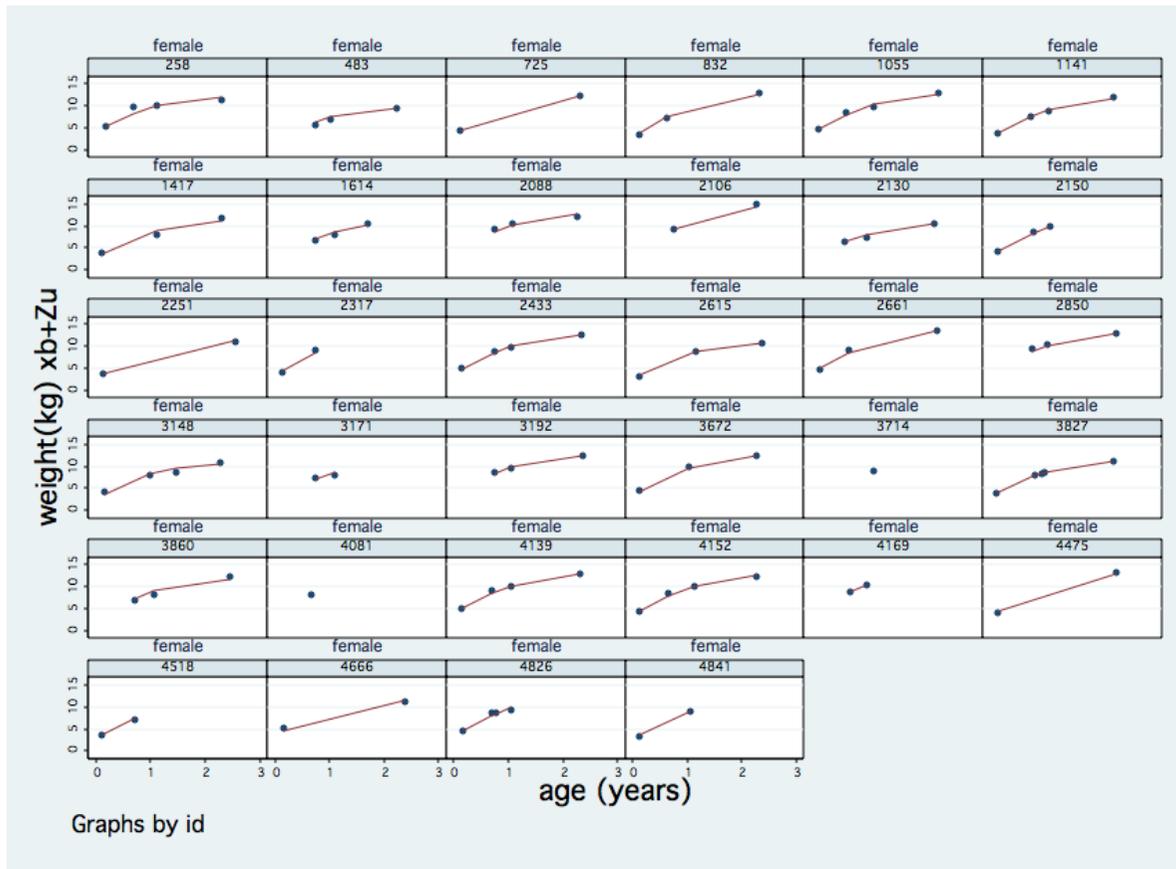
Le assunzioni sono valide?



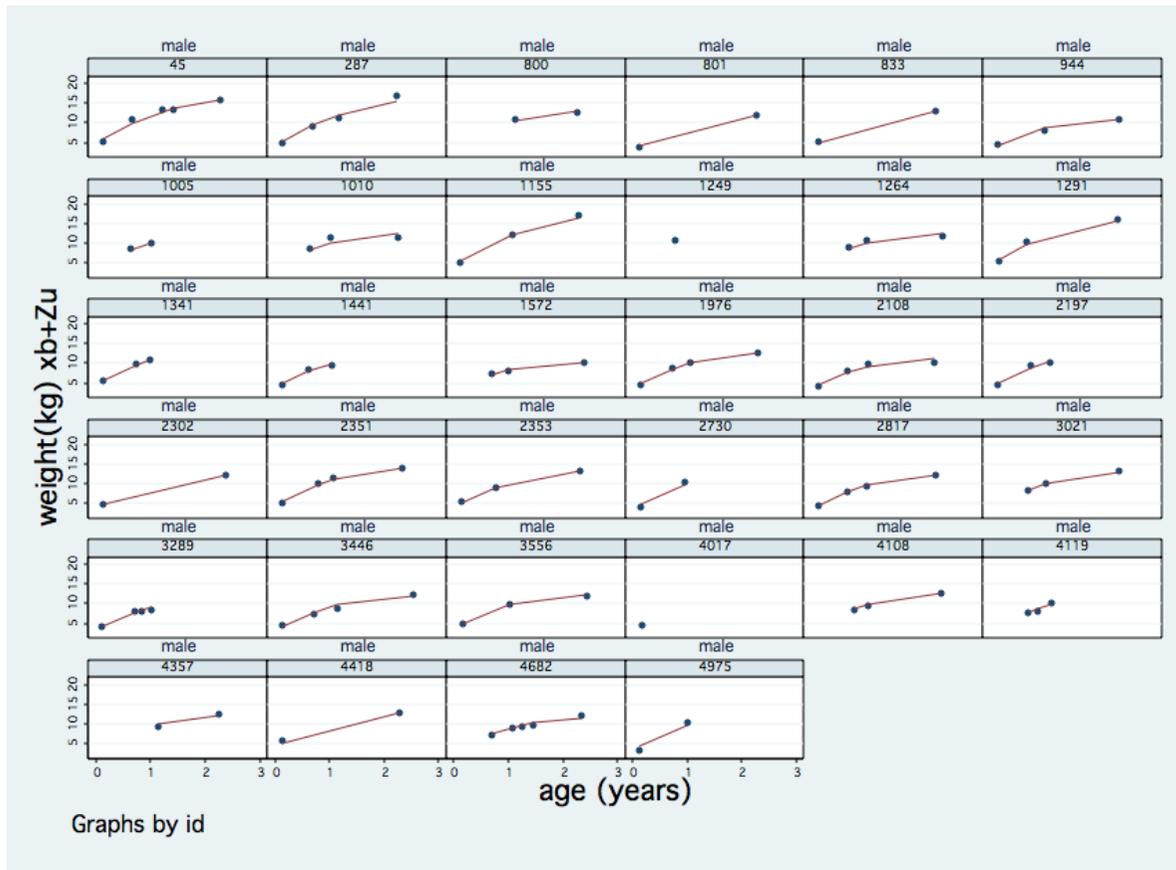
Profileplot valori predetti



Trellis plots feminine



Trellis plots maschi



Cautele 1

- Utilizzare uno slope random senza un intercetto random è (per lo più) privo di significato. Perché?
- Utilizzare un intercetto random senza uno slope random è (per lo più) privo di significato. Perché?
- Modellare molte covariate con slope random è rischioso: questo è un problema generale del "multi-level modeling"

Cautele 2

$$p = \frac{(k + 2)(k + 1)}{2 + 1}$$

- p = numero di parametri testati dal modello
- k = numero di "random slopes" (assumendo 1 "random intercept")

Cautele 3

- È per lo più privo di significato includere uno slope random per una covariata che non si trova a un "livello inferiore" all'intercetto random
- Questo è un problema relativo per i dati longitudinali dove lo slope è il tempo ma è centrale nel "multilevel modeling"
- Non creare mai "gerarchie" che non siano state pre-specificate sulla base di considerazioni biologiche o cliniche

Cautele 4

- Cosa succede se le assunzioni di normalità dei coefficienti random sono violate?
 - Utilizzo intervalli di confidenza robusti (la soluzione più semplice ma non necessariamente la migliore)
 - Modello l'eteroscedasticità ("heteroskedasticity") facendo dipendere la varianza al "livello 1" ed eventualmente al "livello 2" da una o più covariate

ALTRI MODELLI

“Time-lag model”

$$Y_{it} = \beta_0 + \sum_1^J \beta_{1j} X_{ijt-1} + \dots$$

Associa il valore dell'outcome al tempo t a quello del predittore al tempo $t-1$

Quando potrebbe essere utile un modello di questo tipo (e quando è assolutamente inutile)?

$$\begin{bmatrix} Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} + \dots$$

“Change model”

$$(Y_{it} - Y_{it-1}) = \beta_0 + \sum_1^J \beta_{1j} (X_{ijt} - X_{ijt-1}) + \dots$$

Associa la differenza ($t-(t-1)$) dell'outcome alla stessa differenza nel predittore

$$\begin{bmatrix} Y_2 - Y_1 \\ Y_3 - Y_2 \\ Y_4 - Y_3 \\ Y_5 - Y_4 \\ Y_6 - Y_5 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} X_2 - X_1 \\ X_3 - X_2 \\ X_4 - X_3 \\ X_5 - X_4 \\ X_6 - X_5 \end{bmatrix} + \dots$$

Quando potrebbe essere utile un modello di questo tipo?

“Autoregressive model”

$$Y_{it} = \beta_0 + \sum_1^J \beta_{1j} X_{ijt-1} + \beta_{2j} Y_{jt-1} + \dots$$

$$\begin{bmatrix} Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} + \beta_2 \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} + \dots$$

Associa il valore dell'outcome al valore del predittore a $(t-1)$ e al valore dell'outcome a $(t-1)$

Questo è un modello autoregressivo di primo livello

Quando potrebbe essere utile un modello di questo tipo?

Caveat

“In general, it should be realized that it is better to base the choice of a specific longitudinal model on logical considerations instead of statistical ones.

To obtain the most general answer to the question of whether there is a longitudinal relationship between the outcome variable and the predictors and time, the results of several models should be combined.

In practice, however, this almost never happens: a priori the most appropriate model is chosen (usually the ‘standard model’) and only those results are reported.”

Twisk JWR. *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge: Cambridge University Press, 2003, p. 114.

ANALISI DEL CAMBIAMENTO (2 PUNTI)

Analisi del cambiamento

- “Analysis of change”
- Comune nei trial randomizzati controllati (RCT) quando sono utilizzati outcome continui
- (Il metodo è per lo più inadeguato con >2 misurazioni)

Analisi del cambiamento

BMC Medical Research Methodology (2001) 1:6

<http://www.biomedcentral.com/1471-2288/1/6>

Research article

The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study

Andrew J Vickers

Address: Integrative Medicine Service, Biostatistics Service, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue New York, New York 10021, USA

E-mail: vickersa@mskcc.org

Published: 28 June 2001

Received: 28 March 2001

BMC Medical Research Methodology 2001, 1:6

Accepted: 28 June 2001

This article is available from: <http://www.biomedcentral.com/1471-2288/1/6>

© 2001 Vickers, licensee BioMed Central Ltd.

Analisi del cambiamento

- Definiamo:
 - b = valore prima del trattamento ("baseline"/"basale")
 - f = valore dopo il trattamento ("final"/"finale")
 - g = gruppo di trattamento ("group"/"gruppo")

Quattro opzioni a confronto

Denominazione	Come descrivo il risultato
"POST"	f per g
"FRACTION"	$(b-f)/b*100$ per g
"CHANGE"	$(b-f)$ per g
Analisi della covarianza ("ANCOVA")	$f = \beta_0 + \beta_1 b + \beta_2 g$

POST

- Negli studi di simulazione ha la potenza maggiore dopo ANCOVA
- Crea un bias verso il gruppo col valore basale più alto

FRACTION

- Ha la potenza più bassa
- È eccessivamente sensibile a modificazioni della distribuzione del valore basale
- Crea un bias verso il gruppo col valore basale più basso
- Essendo un rapporto, potrebbe non essere normalmente distribuito

CHANGE

- Ha una bassa potenza
- Crea un bias verso il gruppo col valore basale più basso

ANCOVA

- Ha la potenza maggiore
- È il metodo di scelta

*RCT e modelli misti

- Riuscite a immaginare un'“ANCOVA” a effetti random?
- In cosa differirebbe dall' “ANCOVA” tradizionale?
- (Cosa altro potremmo fare con gli “effetti random” nel caso di un RCT?)

Per approfondire l'uso dei modelli a effetti random nei trial clinici:
Brown H, Prescott R. *Applied mixed models in medicine*. NY: Wiley; 2006.

MODULO 3

Modulo 3

- Outcome dicotomici
- "Count outcomes"
- Missing data
- Sample size

OUTCOME DICOTOMICI

Outcome dicotomici

- Metodi tradizionali
- Metodi moderni

Metodi tradizionali

- 2 misurazioni
 - Test di McNemar
 - "Proportion of change"
- > 2 misurazioni

Test Q di Cochran

Test di McNemar

	$t_1 -$	$t_1 +$
$t_0 +$	A	B
$t_0 -$	C	D

t_0 = inizio

t_1 = fine

+ = caratteristica presente

- = caratteristica assente

Testa la differenza tra i cambiamenti in direzioni opposte

Perché si sottrae -1?

$$X^2 = \frac{(A - D - 1)^2}{A + D} (df = 1)$$

“Proportion of change”

	$t_1 -$	$t_1 +$
$t_0 +$	A	B
$t_0 -$	C	D

Qual è la proporzione di soggetti che cambia nel corso dello studio (p_{change})?

Possiamo calcolare una “proportion of increase” e una “proportion of decrease”?

Come calcoliamo l'errore standard (SE) della *differenza* tra due gruppi?

$$SE(p_{change}) = \sqrt{\frac{p_{change} - (1 - p_{change})}{n}}$$

Per approfondire: Fleiss JL, Levin BA, Paik MC. *Statistical methods for rates and proportions*. NY: Wiley; 2003.

Test Q di Cochran

Estensione multivariata del test di McNemar

P_{total} = proportion of *total change*

$$P_{total} = \frac{1}{N(T-1)} \sum_{i=1}^N C_i$$

C_i = numero di cambiamenti a carico dell'individuo i nel tempo T

Qual è il limite evidente di questo test?

Metodi moderni

- Modelli con effetti random
- Equazioni generalizzate di stima

*Il “modello generale” applicato alla distribuzione binomiale

$$\ln \left(\frac{\Pr(Y_{it} = 1)}{1 - \Pr(Y_{it} = 1)} \right) = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t$$

- La distribuzione dell’outcome non è più gaussiana ma binomiale
- Come si chiama il membro di sinistra dell’equazione?
- \ln = logaritmo naturale

Qual è la probabilità dell'evento condizionale al tempo e alle covariate?

$$\Pr(Y_{it} = 1) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t\right)}}$$

- Come per gli outcome continui, β_1 riflette sia la variabilità intra-individuale che quella inter-individuale

Modelli logistici con intercetto random

$$\rho = \frac{\tau^2}{\tau^2 + \pi^2 / 3}$$

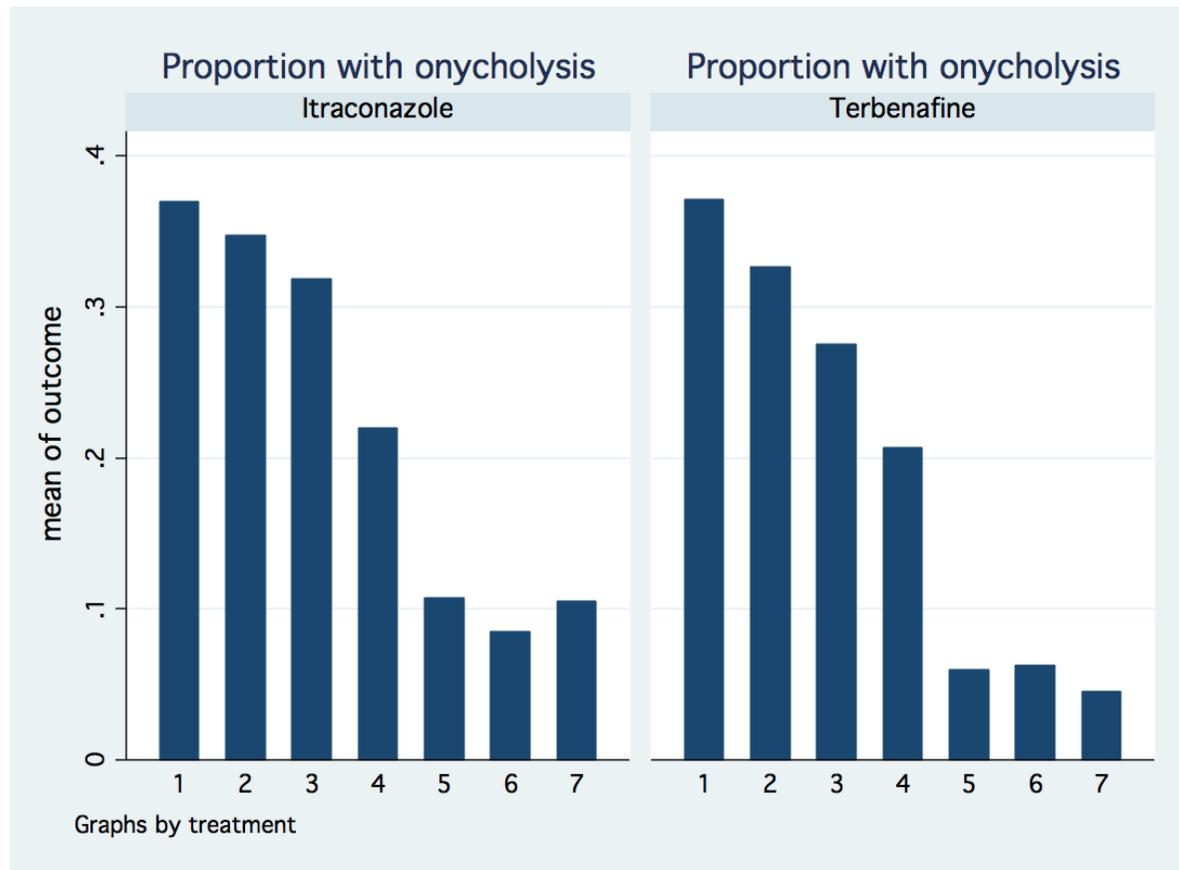
- Similmente ai modelli con outcome continuo, il coefficiente di correlazione intraclassa offre una misura della varianza residua dovuta ai differenti individui
- Perché $\pi^2/3$?

Case study

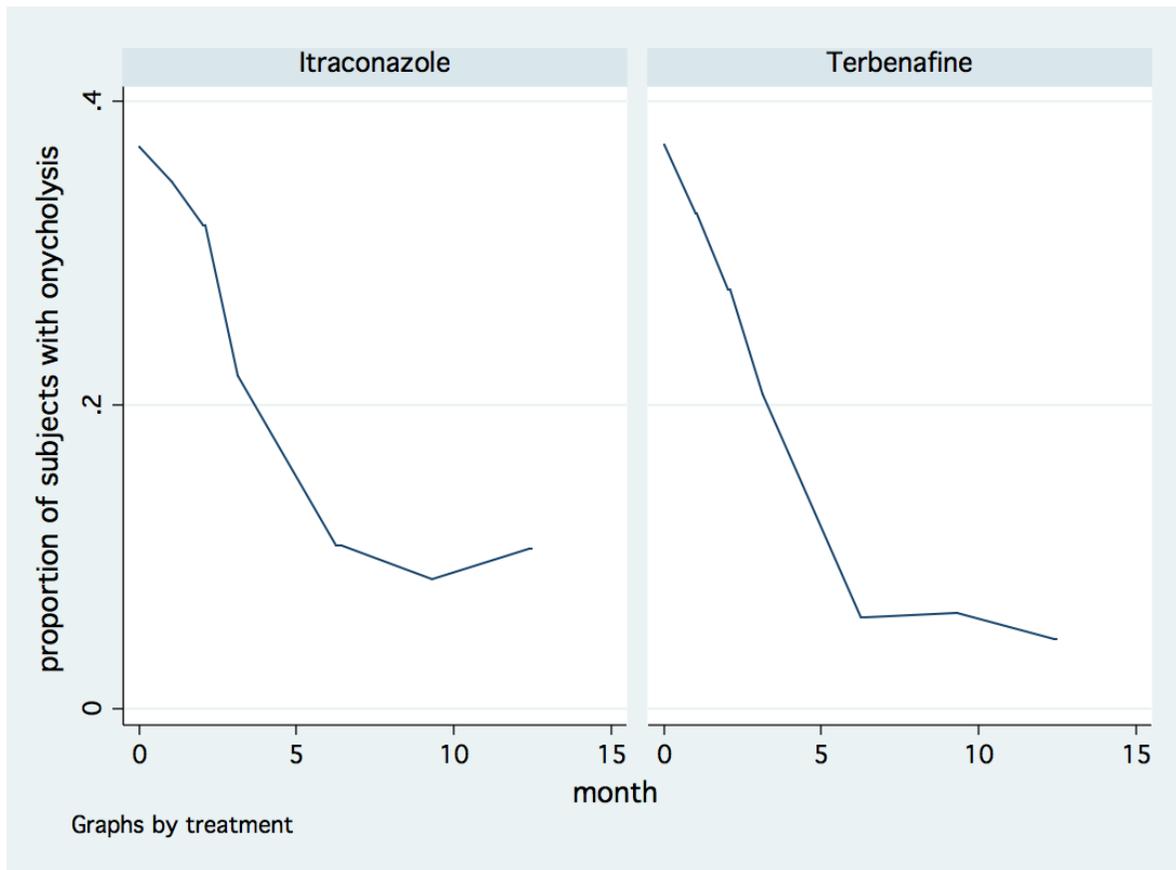
- Trial randomizzato controllato
- 383 pazienti con onicomicosi randomizzati a terbinafina (gruppo sperimentale) o itraconazolo (gruppo di controllo); di 294 è disponibile l'outcome
- Outcome = regressione dell'onicolisi (1= si; 0=no)
- Predittore = trattamento (1 = terbinafina; 0 = itraconazolo)
- Tempo = month = variabile continua "unequally spaced"
- Visit = numero della visita consecutiva (irregolare)

Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modelling using STATA. Stata Press; 2008. www.stata-press.com/data/mlmus2.

Onicolisi in funzione della visita



Onicolisi in funzione del tempo



Regressione logistica "naive"

```
Logistic regression                               Number of obs =      1908
                                                  LR chi2(3)      =      164.47
                                                  Prob > chi2     =      0.0000
Log likelihood = -908.00747                    Pseudo R2      =      0.0830
```

outcome	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
treatment	.9994185	.1560558	-0.00	0.997	.7359277	1.357249
trXmonth	.934988	.0350845	-1.79	0.073	.8686913	1.006344
month	.8434052	.0199212	-7.21	0.000	.8052504	.8833679

Notate l'interazione trXmonth = tempo*trattamento

Quale ipotesi testa e che cosa significa il fatto che non è statisticamente significativa?

Regressione logistica con intercetto random

```

Random-effects logistic regression
Group variable: patient

Random effects u_i ~ Gaussian

Log likelihood = -624.98612

Number of obs      =    1908
Number of groups   =     294

Obs per group: min =     1
                avg =    6.5
                max =     7

Wald chi2(3)      =    151.56
Prob > chi2       =    0.0000
    
```

outcome	OR	Std. Err.	z	P> z	[95% Conf. Interval]	
treatment	.8519413	.4522187	-0.30	0.763	.3010119	2.411213
trXmonth	.8719294	.0593928	-2.01	0.044	.7629576	.9964655
month	.6760442	.0300151	-8.82	0.000	.6197026	.7375082
/lnsig2u	2.813143	.179466			2.461396	3.164889
sigma_u	4.081936	.3662843			3.423618	4.866839
rho	.8351114	.0247125			.7808366	.8780446

Likelihood-ratio test of rho=0: $\chi^2(01) = 566.04$ Prob >= $\chi^2 = 0.000$

Esercizio: calcolate ICC da questi dati

Aggiungiamo uno slope random per il tempo (ha senso?)

Condition Number = **46.89368**

gllamm model

log likelihood = **-540.98242**

outcome	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
treatment	1.008018	.8720214	0.01	0.993	.184972	5.493261
trXmonth	.7363744	.1183632	-1.90	0.057	.5373762	1.009064
month	.2731749	.0664288	-5.34	0.000	.16961	.4399773

Variations and covariances of random effects

***level 2 (patient)

var(1): **100.19591 (27.335599)**

cov(2,1): **-7.1369453 (2.150181)** cor(2,1): **-.58808435**

var(2): **1.4699261 (.44676597)**

Confrontiamo il modello "naive" e quello con intercetto random: cosa notate?

	(1) naive	(2) random intercept
main		
treatment	1.00 [0.74, 1.36]	0.85 [0.30, 2.41]
trXmonth	0.93 [0.87, 1.01]	0.87* [0.76, 1.00]
month	0.84*** [0.81, 0.88]	0.68*** [0.62, 0.74]
N	1908	1908

Exponentiated coefficients; 95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Individuo vs. popolazione

- Il modello logistico standard calcola “population-averaged probabilities” (“marginal”) mentre quello logistico con intercetto random calcola “subject-specific probabilities” (“conditional”)

Diagnostica

- A differenza dei modelli con outcome continuo, i valori stimati degli intercetti random *non* sono utili per la diagnostica
- (Comunque, impostando il modello come GLM abbiamo a disposizione la diagnostica "standard" di questi modelli)

Adesso passiamo alle GEE

```

GEE population-averaged model          Number of obs      =      1908
Group variable:                        patient              Number of groups   =      294
Link:                                   logit               Obs per group: min =      1
Family:                                binomial             avg =              6.5
Correlation:                           exchangeable         max =              7
                                          Wald chi2(3)        =      63.44
Scale parameter:                       1                   Prob > chi2         =      0.0000
  
```

(Std. Err. adjusted for clustering on patient)

outcome	Odds Ratio	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
treatment	1.007207	.2618022	0.03	0.978	.6051549	1.676373
trXmonth	.9252113	.0501514	-1.43	0.152	.8319576	1.028918
month	.8425856	.0253208	-5.70	0.000	.7943911	.893704

Matrice di correlazione

Estimated within-patient correlation matrix R:

	c1	c2	c3	c4	c5	c6	c7
r1	1						
r2	.4217719	1					
r3	.4217719	.4217719	1				
r4	.4217719	.4217719	.4217719	1			
r5	.4217719	.4217719	.4217719	.4217719	1		
r6	.4217719	.4217719	.4217719	.4217719	.4217719	1	
r7	.4217719	.4217719	.4217719	.4217719	.4217719	.4217719	1

"COUNT" OUTCOMES

Distribuzione di Poisson 1

$$\Pr(y | \mu) = \frac{e^{-\mu} \mu^y}{y!}$$

$$\mu = \lambda t$$

Assunzione: gli eventi sono indipendenti e avvengono con la stessa probabilità istantanea o "incidence rate" λ

Distribuzione di Poisson 2

$$\Pr(y_1 \mid \mu = \lambda t) \Pr(y_2 \mid \mu = \lambda t) = \\ \Pr(y_1 + y_2 \mid \mu = \lambda 2t)$$

- Se due (o più) eventi sono Poisson-distribuiti possiamo sommarli ottenendo un intervallo di tempo $2t$
- Pertanto, i dati possono essere aggregati senza perdita di informazione

Distribuzione di Poisson 3

$$\mu_i = E(y_i | x_i) = e^{\beta_1 + \beta_2 x_i} = e^{\beta_1} \times e^{\beta_2 x_i}$$

$$\ln(\mu_i) = \beta_1 + \beta_2 x_i$$

- La relazione con le covariate degli individui (o gruppi) i può essere modellata con un modello log-lineare

Distribuzione di Poisson 4

$$\lambda_i t = e^{\beta_1 + \beta_2 x_i} = e^{\beta_1} \times e^{\beta_2 x_i}$$

- È la precedente riscritta

Distribuzione di Poisson 5

$$\frac{\lambda_i}{\lambda'_i} = \frac{\lambda_i t}{\lambda'_i t} = \frac{e^{\beta_1} \times e^{\beta_2 x_i}}{e^{\beta_1} \times e^{\beta_2 x'_i}} = e^{\beta_2 (x_i - x'_i)} = e^{\beta_2 (x_i - x'_i)}$$

- Il valore del coefficiente di regressione “esponenziato” (e^β) corrisponde a un “incidence rate ratio”

Case study

- Trial randomizzato controllato
- 59 soggetti epilettici
- Outcome = numero delle crisi epilettiche durante le 2 settimane prima di ognuna di 4 visite cliniche

Skrondal A et al. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Boca Raton: Chapman & Hall; 2004.

*Case study

- Predittori
 - Intraindividuali
 - Lbas = logaritmo di 1/4 del numero delle crisi epilettiche nelle 8 settimane prima del trial
 - Treat = progabide (1) vs. placebo (0)
 - Lbas_Treat = interazione tra Lbas e Treat
 - Lage = logaritmo dell'età
 - Interindividuali
 - V4 = visita 4 (1 = sì, 0 = no)

Il modello e le sue assunzioni

$$y_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\ln(\mu_{ij}) = \beta_{0j} + \beta_{1j} * Lbas + \beta_{2j} * Treat + \\ \beta_{3j} * Lbas_Treat + \beta_{4j} * Lage_j + \beta_{5ij} * V4 \\ + u_j$$

$$u_j \sim N(0, \sigma^2)$$

Case study

Condition Number = **9.3176679**

gllamm model

log likelihood = **-665.29074**

y	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
lbas	2.42161	.3177909	6.74	0.000	1.872406	3.131904
treat	.3933572	.1576711	-2.33	0.020	.1793066	.862935
lbas_treat	1.402505	.2851832	1.66	0.096	.9415046	2.089231
lage	1.622936	.5636091	1.39	0.163	.8216718	3.205564
v4	.8512179	.0464559	-2.95	0.003	.7648667	.9473178

Variances and covariances of random effects

***level 2 (subj)

var(1): **.25282624 (.05895587)**

Case study

Condition Number = **9.3176679**

gllamm model

log likelihood = **-665.29074**

Robust standard errors

y	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
lbas	2.42161	.2710642	7.90	0.000	1.944574	3.015671
treat	.3933572	.1576186	-2.33	0.020	.1793535	.8627093
lbas_treat	1.402505	.2813653	1.69	0.092	.9465412	2.078114
lage	1.622936	.4915371	1.60	0.110	.8963938	2.938353
v4	.8512179	.0563037	-2.44	0.015	.7477186	.9690436

Variances and covariances of random effects

***level 2 (subj)

var(1): **.25282624 (.06247829)**

MISSING DATA

Missing data

- Non tutti gli N individui hanno le misure di interesse per tutti i tempi T dello studio
- I missing data possono essere "intermittenti" o "permanenti"
- (In genere, oltre a un certo numero di missing data consecutivi, si definisce un "drop-out")

Classificazione di Rubin

- Cominciamo col caratterizzare la distribuzione dei missing data o "DOM" (distribution of missingness):
 - "Missing completely at random" (MCAR): DOM non dipende né dalle covariate né dall'outcome. Esempio?
 - "Missing at random" (MAR): DOM può dipendere dalle covariate e dall'outcome. Esempio?
 - "Missing not at random" (MNAR): ogni violazione di MAR. Esempio?

Perché classifichiamo la DOM?

- Lo scopo non è modellare il pattern di missingness (raramente abbiamo le informazioni per farlo!)
- Lo scopo è chiarire le condizioni in cui la missingness essere ignorata ("ignorable")
- Nella discussione che segue utilizzeremo un approccio frequentista (le conclusioni sono diverse con un approccio bayesiano)

Perché classifichiamo la DOM?

- Possiamo spesso testare MCAR e MAR
- Non possiamo mai testare NMAR
- Un pattern MCAR può essere ignorato, non gli altri

Metodi per missing data

- Case deletion
- Mean or median of series
- Hot deck
- Cross-sectional linear regression
- Last observation carried forward (LOCF)
- Linear interpolation
- Longitudinal linear regression
- *Likelihood*
- *Multiple imputation*
- *Reweighting*

Metodi per missing data

- Le uniche tecniche efficaci sono le ultime 3:
 - Likelihood
 - Multiple imputation
 - Reweighting
- Le altre le consideriamo unicamente per ragioni didattiche (purtroppo sono ancora di uso comune nella letteratura biomedica)

Case deletion

- Elimino gli individui con informazioni incomplete
- Può funzionare sotto MCAR ma se la missingness è alta introduce bias nella stima
- Con uno studio longitudinale, il rischio di eliminare molti dati è alto perché variabili differenti possono essere "missing" a tempi differenti
- Si tratta comunque di una procedura inefficiente

Mean or median of series

- Prendo la media o la media della "cross-section" (uno studio longitudinale può essere pensato come costituito da tante "cross-sections")
- Difficile da giustificare teoricamente
- Meno robusta della "case deletion" se la DOM è MCAR!

Hot deck

- Prendo valore medio (o mediano o random) di un gruppo comparabile di soggetti
- Il numero di soggetti può variare da 1 a tutta la popolazione disponibile
- È un metodo migliore del precedente perché preserva qualche aspetto della distribuzione ma presenta ancora parecchi limiti

Cross-sectional linear regression

- Utilizzo il risultato di una regressione lineare per predire l'outcome Y dalle variabili disponibili sulla "cross section"
- Può funzionare se i dati sono MAR e utilizzo un valore random della distribuzione condizionale di Y su X
- A questo punto, è molto più semplice fare una multiple imputation, che ha molti più vantaggi

Last observation carried forward (LOCF)

- Utilizzo l'ultimo valore disponibile della misurazione
- Nessuna giustificazione teoretica
- Giustificazione pratica inesistente dal momento che esistono tecniche di gran lunga superiori
- Da non usare mai!

Longitudinal linear regression

- Utilizzo la regressione lineare per interpolare un valore tra più punti separati nel tempo
- Evoluzione della cross-sectional linear regression
- Ancora problematica

*Maximum likelihood

- *Se i dati sono MAR, il likelihood corretto può essere ottenuto dalla distribuzione dei dati osservati*
- Tutte le tecniche che utilizzano la tecnica MLE (o quasi-likelihood) possono utilizzare questa modalità di gestione dei missing data
- Per i modelli a coefficienti random, i dati devono essere MAR
- Per le GEE, i dati devono essere invece MCAR (anche se la letteratura al riguardo è controversa)

Maximum likelihood

- Tutti i modelli che abbiamo considerato in questo corso hanno utilizzato il metodo MLE per gestire “internamente” i missing data
- Torniamo indietro a verificare se le assunzioni di una DOM MAR per modelli con coefficienti random e di una DOM MCAR per GEE erano accettabili

Multiple imputation

- La tecnica di scelta in molti contesti
- Nettamente preferibile alla "single imputation" perché aggiunge un' "incertezza" (= variabilità) alla stima del dato

Multiple imputation

- La procedura è la seguente
 - Genera un numero M di dataset completi utilizzando un modello di regressione del dato "missing" vs. altre variabili disponibili
 - Effettua l'analisi su ciascuno degli M dataset
 - Combina gli M dataset per fornire una sola statistica

Multiple imputation

$$Avg = \frac{\sum_{i=1}^M b}{M}$$

$$Var_w = \frac{\sum_{i=1}^M var_i}{M}$$

$$Var_b = \frac{\sum_{i=1}^M (b_i - \bar{b})^2}{M}$$

$$Var = Var_w + \frac{M+1}{M} Var_b$$

Il metodo più semplice di calcolo

Avg = media; b = parametro di interesse; M = numero di dataset; Var_w = within-imputation variance; Var_b = ; between imputation variance; = valore singolo del parametro di interesse.

Non vi sarà sfuggito che stiamo assumendo una distribuzione normale multivarata!

Quanti dataset M ci servono?

- Per quanto sorprendente possa sembrare, gli studi di simulazione dimostrano *chiaramente* che 5 dataset sono sufficienti per ottenere una stima accurata della media e della varianza del parametro di interesse
- Splendido, ma come la mettiamo con l'assunzione della normalità multivariata?

E la normalità multivariata?

- Ancora più sorprendentemente, la MI è molto robusta alle violazioni della normalità, anche in piccoli campioni
- (Saremo nati con la camicia?)
- In ogni caso, esistono trattamenti matematici appositi per affrontare le deviazioni dalla normalità (questa è un'area di ricerca in rapidissimo sviluppo)

E la normalità multivariata?

“Real data rarely conform to normality.

Some might hesitate to use a normal imputation model, fearing that it may distort the distributions of nonnormal variables.

Many tricks are available to preserve distributional shape.”

Schafer JL, *Psychol Methods* 2002, 7:147

Reweighting

$$w_{ij}^{-1} = \pi_{ij} = \Pr(r_{ij} = 1 \mid data)$$

- “Weighted estimating equations” (WEE)
- L’idea è di modellare la nostra propensione ad eliminare un dato in funzione dell’informazione (“data”) che abbiamo e che sembra correlata alla missingness
- Campo in continua evoluzione

E sei dati sono MNAR?

- MNAR è spesso un problema nei trial clinici, dove i partecipanti potrebbero uscire dallo studio per ragioni correlate all'outcome di interesse

*E sei dati sono MNAR?

- Ho due opzioni principali
 - “Selection model”: specifico una distribuzione per i dati completi e propongo una forma per la dipendenza della DOM dai dati.
 - “Pattern mixture model”: classifico gli individui per pattern di missingness e descrivo i dati all’interno dei gruppi di missingness

Ma non è tutto ipotetico?

- Esatto ! Stiamo “modellando” eventi che non conosciamo affatto e dobbiamo procedere con estrema cautela
- L'unico uso “sensato” di questi metodi è un'analisi di sensibilità dei risultati del trial clinico. Il “pattern mixture model” è attualmente quello di scelta
- In generale, non è consigliabile fare assunzioni sui dati MNAR al di fuori dei trial clinici. Perché?

In conclusione

“Although other procedures are occasionally useful, we recommend that researchers apply likelihood-based procedures, where available, or the parametric MI described in this article, which are appropriate under general MAR conditions.

As MNAR methods are incorporated into mainstream software, they, too, become attractive in certain circumstances, particularly for sensitivity analysis.

Until then, ML under MAR assumption represent the practical state of the art.”

Schafer JL, *Psychol Methods* 2002, 7:147

Strategie per l'analisi (cenni)

- Cominciare sempre descrivendo il pattern di missingness utilizzando tecniche esplorative
 - Regressione logistica
 - Cluster analysis

Per approfondire: Harrell FE. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer; 2001. (Attesa nuova edizione).

Lecture consigliate

- Cominciate da
 - Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychol Methods* 2002;7:147 - livello intermedio.
 - Allison PD, *Missing Data*. Thousand Oaks: Sage; 2005. Livello principiante/intermedio.
- Proseguite con
 - Little RJA et al. *Statistical analysis with missing data*. New York: Wiley; 2002 - livello intermedio/avanzato.
 - Carpenter JR et al. *Missing data in randomised controlled trials - a practical guide*. Birmingham: National Co-ordinating Centre for Research methodology; 2007 - livello intermedio/avanzato. Disponibile gratuitamente previa registrazione a missingdata.com.

SAMPLE SIZE PER STUDI LONGITUDINALI

Sample size e studi longitudinali

- Formule e algoritmi sono ben sviluppati solo per gli RCT
- Per gli studi osservazionali, è spesso necessario effettuare “simulazioni” dello scenario di ricerca
- Partirò col considerare le formule per un RCT longitudinale
- Quindi, illustrerò un metodo “ibrido” per il calcolo del sample size per uno studio osservazionale longitudinale

RCT con 2 gruppi e 2 misurazioni
di outcome continuo

$$N = \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta/2)} \right)^2 \sigma^2 (r + 1)}{v^2 r}$$

N = numero di soggetti

$Z_{(1-\alpha/2)}$ = (1- α /2) percentile (errore tipo I)

$Z_{(1-\alpha\beta/2)}$ = (1- β /2) percentile (errore tipo II)

r = rapporto dei soggetti tra i gruppi

σ^2 = varianza dell'outcome

v = differenza dell'outcome tra i due gruppi

RCT con 2 gruppi e >2 misurazioni
di outcome continuo

$$N = \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta/2)} \right)^2 \sigma^2 (r + 1) [1 + (T - 1)\rho]}{v^2 r T}$$

T = numero di misurazioni ripetute

ρ = *coefficiente di correlazione per misure ripetute*

*RCT con 2 gruppi e 2 misurazioni
di outcome dicotomico

$$N = \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta/2)} \right)^2 p(1-p)(r+1)}{(p_1 - p_0)^2 r}$$

$$p = \frac{p_1 + rp_0}{1+r}$$

p_1 = numero di "casi" nel gruppo sperimentale

p_0 = numero di "casi" nel gruppo di controllo

p = media "pesata" dei casi tra i due gruppi

RCT con 2 gruppi e >2 misurazioni
di outcome dicotomico

$$N = \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta/2)} \right)^2 p(1-p)(r+1)[1 + (T-1)\rho]}{(p_1 - p_0)^2 rT}$$

T = numero di misurazioni ripetute

ρ = *coefficiente di correlazione per misure ripetute*

*Case study

- Outcome continuo
- 2 gruppi di dimensioni uguali
- 3 misurazioni con nessun drop-out
- $\rho = 0.5$ costante
- $v = 0.25$
- $\sigma^2 = 1$
- $r = 1$
- $\alpha = 0.05$
- $\beta = 0.90$

*Risolviamo l'equazione

$$N = \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta/2)} \right)^2 \sigma^2 (r + 1) [1 + (T - 1)\rho]}{v^2 r T}$$

$$N = \frac{(1.96 + 1.28)^2 \times 1 \times (1 + 1) \times [1 + (3 - 1) \times 0.5]}{0.25^2 \times 1 \times 3} = 224$$

Simulazione scenario

```
. input g t

           g      t
1. 0 1
2. 0 2
3. 0 3
4. 1 1
5. 1 2
6. 1 3
7. end

. expand 224
(1338 observations created)

. bysort t: gen id=_n

.
. * creo una variabile random y compresa tra 0 e 1
. gen y = uniform()

.
. *creo la matrice di correlazione dello studio
. mat R = J(3,3,0.5) + I(3)*(1-0.5)

. mat list R

symmetric R[3,3]
      c1  c2  c3
r1   1
r2  .5  1
r3  .5  .5  1
```

Simulazione scenario

```
. *utilizzo un'equazione di stima generalizzata con la matrice che ho creato
. *per stimare y da g
. xtgee y g, i(id) t(t) corr(fixed R) scale(1)
```

```
Iteration 1: tolerance = .01060198
Iteration 2: tolerance = 0
```

```
GEE population-averaged model
Group and time vars:          id t      Number of obs      =      1344
Link:                          identity  Number of groups   =      448
Family:                         Gaussian  Obs per group: min =      3
Correlation:                    fixed (specified)  avg =      3.0
Scale parameter:                1          max =      3
                                Wald chi2(1)       =      0.01
                                Prob > chi2        =      0.9347
```

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
g	-.0036794	.0449013	-0.08	0.935	-.0916844	.0843256
_cons	.5032601	.0446333	11.28	0.000	.4157805	.5907397

Simulazione scenario

```
. *adesso utilizzo il valore stimato di g  
. *come deviazione standard  
. *per calcolare il sample size  
. *di uno studio con 1 gruppo  
. sampsi 0 0.25, alpha(0.05) n1(1) sd1(.0014784) onesample
```

Estimated power for one-sample comparison of mean
to hypothesized value

Test Ho: $m = 0$, where m is the mean in the population

Assumptions:

```
alpha = 0.0500 (two-sided)  
alternative m = .25  
sd = .001478  
sample size n = 1
```

Estimated power:

```
power = 1.0000
```

GRAZIE!