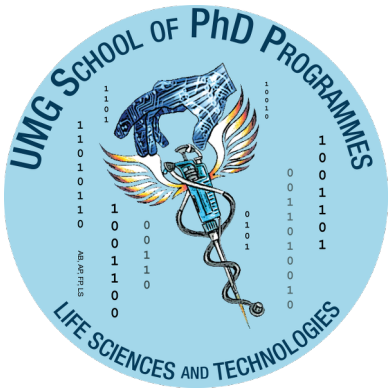
SOCIETA' DEI MATEMATICI
E NATURALISTI DI MODENA
www.socnatmatmo.unimore.it



Massimo Borelli, Ph.D.

il dataset `doublekinetics` analizzato con R

24 marzo 2017

Riproduzione pittorica di foto d'epoca a cura di Caterina Stella.

Giovanni Canestrini, nato a Revò (Tn) nel 1835 e defunto a Padova nel 1900, fu fondatore nel 1865 della Società dei Naturalisti e Matematici di Modena. Nata con lo scopo di promuovere lo studio delle Scienze Naturali e Matematiche, di favorire i legami fra studiosi, ricercatori, collezionisti e diffondere cultura scientifica a tutti i livelli per una equilibrata crescita civile, la Società dei Naturalisti e Matematici di Modena ha accolto nelle sua fila, tra gli altri, Charles Darwin, Louis Pasteur e Thomas Henry Huxley.

Indice

caro Massimo,
io lo vedo chiaramente; ma come posso affermare 'in senso statistico' che la cinetica del mio farmaco esibisce due comportamenti molto diversi? E come faccio a dire che questa differenza è significativa (ammesso che lo sia ..)?

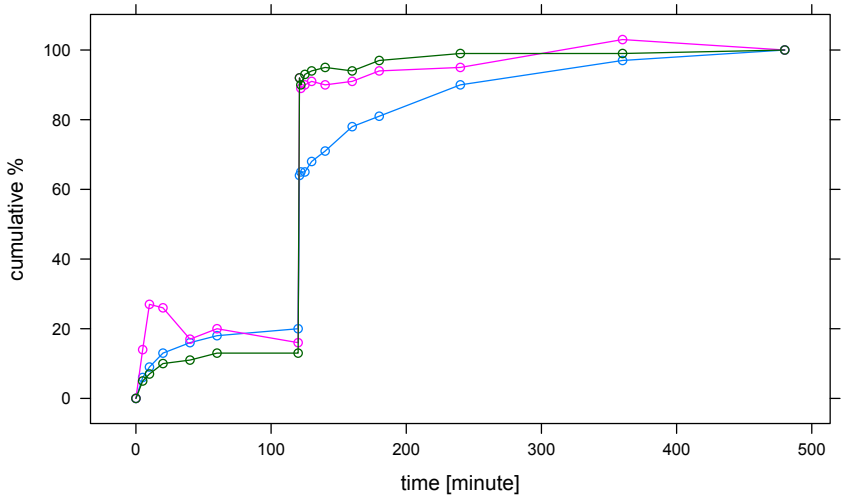


Figura 1: una 'doppia' farmacocinetica

1 Motivazione dello studio

L'assorbimento di un farmaco è strettamente connesso alla sua solubilità, e quest'ultima dipende da diversi fattori; fra questi, l'affinità fra la molecola di interesse e il solvente in cui si vuole solubilizzarla. Ideale sarebbe riuscire a trovare una formulazione che sia in grado di variare la solubilità in relazione al pH del solvente in cui essa si trova.
Nello studio in questione, condotto dal dottor Fabio Tentor presso il Dipartimento di micro- e nanotecnologie dell'Universtià di Lyngby (Danimarca), accade proprio questo: aumentando il pH, la solubilità della molecola cambia drasticamente, esibendo una cinetica che potremmo definire di tipo 'burst'.
Questa peculiarità potrebbe dare consistenti vantaggi nel campo della farmaceutica: ad esempio, in riferimento ad un dosaggio orale, risulterebbe possibile sfruttare la fisiologica variazione di pH fra lo stomaco e l'intestino al fine di ottenere un miglior assorbimento del farmaco a livello duodenale o digiunale.

2 Struttura del dataset `doublekinetics`

Si tratta di un dataset di 51 osservazioni di 4 variabili, in formato *.csv* (il carattere separatore è infatti una virgola): `time` è il tempo misurato in minuto di svolgimento dall'esperimento. Si tratta, dunque, di un esperimento **longitudinale**. La variabile `sample` indica che l'esperimento è stato condotto in triplicato, e dunque siamo in presenza di **misure ripetute**. La variabile `pH` è un fattore a due livelli, `first` e `second`, che modella le due diverse condizioni di assorbimento nello stomaco e nell'intestino. Infine la risposta `cumulative` è una misura, espressa in percentuale, dell'assorbimento del farmaco.

```
www = "http://www.biostatisticaumg.it/dataset/doublekinetics.csv"
doublekinetics = read.csv(www, header = TRUE)
# doublekinetics = read.csv(file.choose(), header = TRUE)
attach(doublekinetics)
head(doublekinetics)
tail(doublekinetics)
str(doublekinetics)
```

I comandi `head` e `tail` ci consentono di visualizzare le prime e le ultime righe del dataset:

	time	sample	pH	cumulative
1	0	A	first	0
2	5	A	first	6
3	10	A	first	9
4	20	A	first	13
5	40	A	first	16
6	60	A	first	18
	..	..	..	..
46	140	C	second	95
47	160	C	second	94
48	180	C	second	97
49	240	C	second	99
50	360	C	second	99
51	480	C	second	100

Il dataset si presenta nel cosiddetto *long format*, particolarmente gradito dai software. Infatti le 51 righe del dataset, che raccolgono 3 'campioni', elencano $51/3 = 17$ misure

`cumulative` raccolte in 17 differenti occasioni temporali (rispettivamente a 0, 5, 10, 20, 40, 60, 120, 121, 122, 125, 130, 140, 160, 180, 240, 360 e 480 minuti). In realtà, gli studiosi trovano comodo raccogliere i dati nel cosiddetto *short format*: nell'immagine sottostante vediamo proprio come Fabio aveva originariamente codificato i dati, con il colore giallo/verde che rappresentava l'informazione `pH`, `first` e `second`.

	Sample A		Sample B	Sample C
2	time (min)	cumulative% vs Max [C]	cumulative% vs Max [C]	cumulative% vs Max [C]
3	0	0	0	0
4	5	6	14	5
5	10	9	27	7
6	20	13	26	10
7	40	16	17	11
8	60	18	20	13
9	120	20	16	13
10	121	64	92	92
11	122	65	89	90
12	125	65	90	93
13	130	68	91	94
14	140	71	90	95
15	160	78	91	94
16	180	81	94	97
17	240	90	95	99
18	360	97	103	99
19	480	100	100	100

Figura 2: Il dataset `doublekinetics` nel suo formato *short*

3 Visualizziamo il dataset `doublekinetics`

Al solito, vi sono varie possibilità per visualizzare i dati del nostro dataset. Il grafico della Figura 1 è stato ottenuto con i seguenti comandi:

```
library(lattice)
xyplot(cumulative ~ time , type = "b", groups = sample ,
xlab = "time [minute]", ylab = "cumulative %")
```

4 Modelliamo la risposta `cumulative`

4.1 Linearizzare la risposta

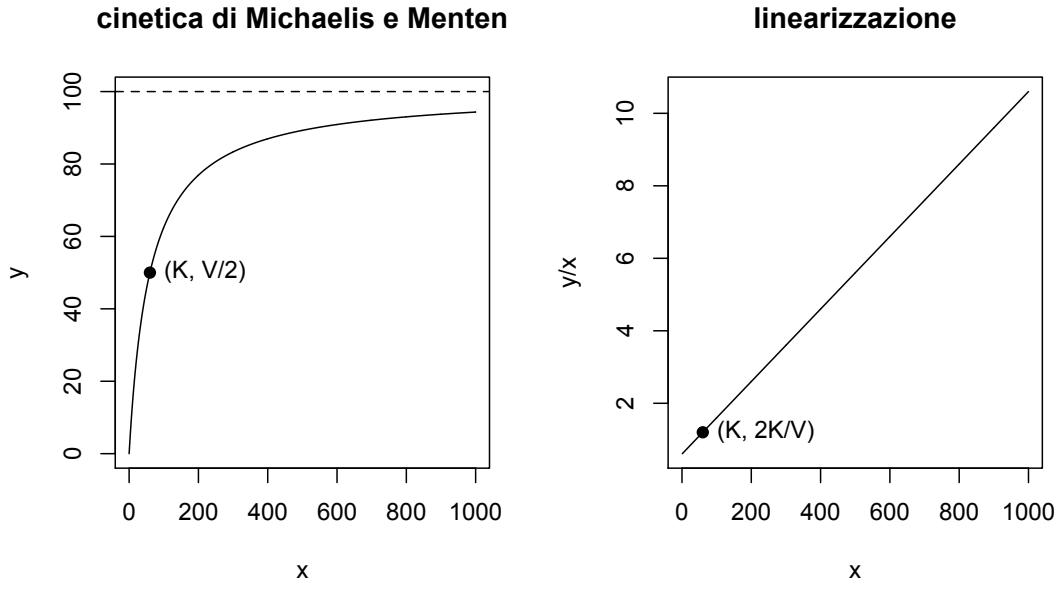


Figura 3: La cinetica di Michaelis e Menten e la sua trasformazione lineare

Innanzitutto, ricordiamo che la cinetica di Michaelis e Menten[5] ha un andamento iperbolico che si può scrivere nella forma:

$$y = \frac{Vx}{K + x}$$

Nel pannello di sinistra della Figura 3, l'asintoto orizzontale tratteggiato ha equazione $y = V > 0$ (nel nostro caso $V = 100\% = 1$), mentre la cosidetta *costante di semisaturazione* rappresenta il tempo $K > 0$ in cui la curva raggiunge la quota $V/2$. È noto[4] che quando $x, y > 0$ tale legge si può trasformare in una relazione lineare, mediante i seguenti passaggi algebrici:

$$\frac{1}{y} = \frac{K + x}{Vx}$$

$$\frac{x}{y} = \frac{K + x}{V}$$

$$\frac{x}{y} = \left(\frac{K}{V}\right) + \left(\frac{1}{V}\right)x$$

$$\frac{x}{y} = a + bx$$

Dunque, la legge di Michaelis e Menten se viene raffigurata su una piano cartesiano di ascissa x ed ordinata x/y diventa una retta che ha pendenza $b = \frac{1}{V}$ ed intercetta $a = \frac{K}{V}$, come vediamo nel pannello di destra della Figura 3.

Pertanto, trasformando la risposta `cumulative` in questo modo:

```
linearized = time / cumulative
```

ecco che il time plot originario, rappresentato nel pannello di sinistra della Figura 4 (in giallo i punti con `Ph first`, in verde `Ph second`, come in Figura 2) che aveva un comportamento non lineare, più difficile da studiare, diviene nel pannello di destra immediatamente trattabile con un **modello statistico di tipo lineare**.

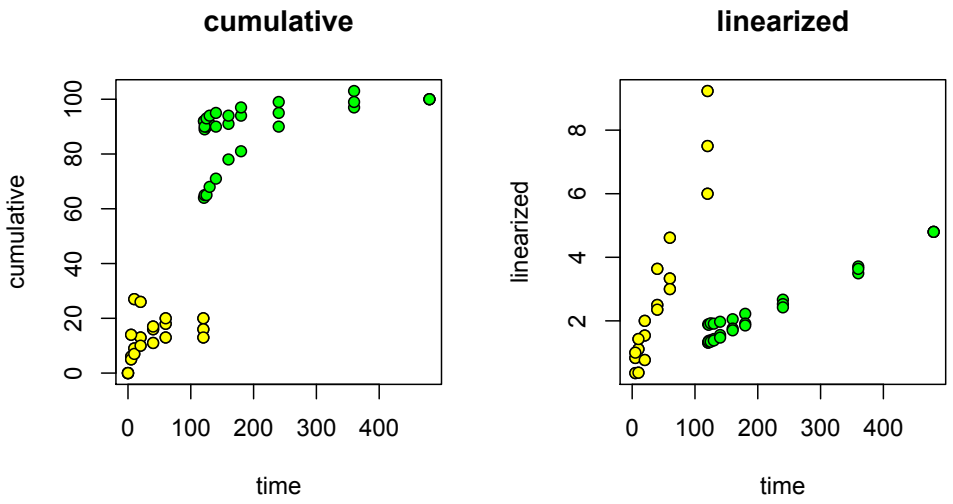


Figura 4: I valori sperimentali del dataset e la loro trasformazione lineare

4.2 Gestire i dati correlati

È sempre buona norma cercare di avere a che fare con ascisse ed ordinate che abbiano il medesimo ordine di grandezza, per evitare che vi siano problemi di convergenza negli algoritmi iterativi utilizzati dal software R. A tale proposito, ci serviamo della funzione `scale` che standardizza le variabili (ossia le trasla nello zero rispetto alla media, e le 'riduce' ad uno rispetto alla deviazione standard):

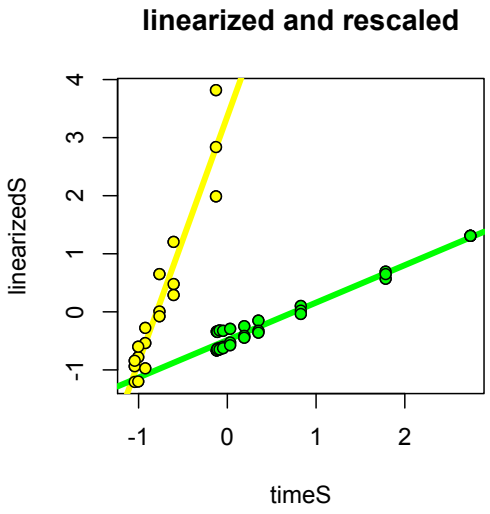
```
linearizedS = scale(linearized)
timeS = scale(time)
```

4.2.1 Un errore da non commettere

Siccome l'esperimento longitudinale è stato anche condotto in triplicato (i.e. **repeated measures**), il dataset non è composto da dati tra loro indipendenti ma **correlati** (nel senso attribuito a questo termine da [6]). Pertanto NON è affatto corretto utilizzare un modello Ancova (i.e. un modello lineare ad **effetti fissi**) per descrivere i dati:

```
sbagliato = lm(linearizedS ~ timeS * pH)
summary(sbagliato)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.3667	0.1700	19.81	0.0000
timeS	4.2311	0.2103	20.12	0.0000
pHsecond	-3.8533	0.1801	-21.39	0.0000
timeS:pHsecond	-3.5858	0.2174	-16.49	0.0000



Intendiamoci: non sono le *Estimate* del modello `sbagliato` ad essere errate. Viceversa, sono gli *Standard Error* ad esserlo, perché essi sono determinati nell'ipotesi che le 51 osservazioni di `linearizedS` (ossia di `cumulative`) siano tra loro indipendenti. Conseguentemente, anche i quozienti *t value* = *Estimate*/*Standard Error* risultano errati.

Infine, un ulteriore errore proviene dal fatto che in presenza di misure ripetute la variabile aleatoria che modella i consuntivi t della terza colonna NON è la classica distribuzione di Student[3, 1], e quindi i p *value* riportati nella quarta colonna sono (doppiamente ..) errati.

4.2.2 Il modo corretto di procedere

Correttamente, dobbiamo condurre un'analisi che coinvolga sia la presenza di **effetti fissi** (il tempo `time` e il fattore `pH`) e di **effetti casuali** (la variabile di *cluster* `sample` che tiene conto delle *pseudorepliche* [2]). Ci serviamo perciò del pacchetto `lme4` [1] per individuare il **modello lineare ad effetti misti** adeguato a descrivere il dataset.

Il modello massimale `mixed0`. Il modello lineare massimale si ottiene utilizzando questa sintassi:

```
library(lme4)
mixed0 = lmer(linearizedS ~ timeS * pH + (timeS | sample))
summary(mixed0)
```

Discutiamo innanzitutto la sezione dedicata agli effetti fissi:

	Estimate	Std. Error	t value
(Intercept)	3.37	0.17	20.06
timeS	4.23	0.19	21.76
pHsecond	-3.85	0.16	-24.03
timeS:pHsecond	-3.59	0.19	-18.53

Come dicevamo poco fa, le *Estimate* che otteniamo sono le stesse di prima, mentre gli *Standard Error* ed i t *value* sono mutati. Il termine di interazione `timeS:pHsecond` ha un consuntivo t *value* $= -18.53$ che è senza alcun dubbio significativo per qualsiasi variabile aleatoria che esista al mondo. In altri termini, le rette di regressione gialla e verde hanno pendenze significativamente diverse, ed è proprio questa la risposta che andavamo cercando:

> *.. E come faccio a dire che questa differenza è significativa (ammesso che lo sia ..)?*

Tuttavia, per decidere se `mixed0` sia il modello minimale adeguato, andiamo anche ad esaminare la sezione degli effetti casuali:

	grp	var1	var2	vcov	sdcor
1	sample	(Intercept)		0.02	0.13
2	sample	timeS		0.01	0.09
3	sample	(Intercept)	timeS	-0.01	-1.00
4	Residual			0.06	0.25

La colonna *sdcor* ci dice che la matrice di correlazione degli effetti casuali è:

$$\begin{pmatrix} 0.13 & -1 \\ -1 & 0.09 \end{pmatrix}$$

Dunque la correlazione tra la perturbazione casuale dell'intercetta e quella della pendenza è altissima. Proviamo dunque a semplificare il modello.

Il modello minimale adeguato `mixed1`. Il modello lineare seguente:

```
mixed1 = lmer(linearizedS ~ timeS * pH + (1 | sample))
summary(mixed1)
```

presenta degli effetti fissi del tutto paragonabili al modello massimale, le differenze sono davvero lievi:

	Estimate	Std. Error	t value
(Intercept)	3.37	0.17	19.53
timeS	4.23	0.20	21.42
pHsecond	-3.85	0.17	-22.77
timeS:pHsecond	-3.59	0.20	-17.55

Questa invece è la sezione degli effetti casuali:

	grp	var1	var2	vcov	sdcor
1	sample	(Intercept)		0.01	0.11
2	Residual			0.07	0.26

5 Interpretazione del modello minimale adeguato

Dall'output del modello `mixed1` possiamo dire che, riferendoci alla Figura di pagina 8, che la retta di regressione gialla e quella verde hanno rispettivamente equazione:

$$y = (3.37 + \beta) + 4.23 \cdot x + \varepsilon$$

$$y = (-0.48 + \beta) + 0.65 \cdot x + \varepsilon$$

essendo β una perturbazione casuale dell'intercetta, distribuita normalmente con media nulla e deviazione standard 0.11, ed essendo ε una perturbazione casuale residua, della medesima natura gaussiana ma con deviazione standard 0.26.

Riferimenti bibliografici

[1] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. R package version 1.1-7.

[2] Michael J Crawley. *Statistics: an introduction using R*. John Wiley & Sons, 2005.

[3] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2005.

[4] Sergio Invernizzi. *Matematica nelle Scienze Naturali*. Edizioni Goliardiche, Trieste, 1996.

[5] Leonor Michaelis and Maud L Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[6] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2000.