Università degli Studi di Trieste

Corso di Laurea Magistrale a Ciclo Unico
LM-41 Medicina e Chirurgia

Dispensa Didattica

810ME
Metodologie di analisi dei dati biomedici

www.dmi.units.it/borelli

Autore
Massimo Borelli, Ph.D.

Anno Accademico 2015 – 2016

Indice

1 Premessa

Questa dispensa, ovviamente, non è un libro di testo. Abbiamo semplicemente raccolto e/o ampliato, in maniera non troppo disordinata, alcuni contenuti delle lezioni del corso. Nelle lezioni abbiamo dato spazio a considerazioni di tipo teorico e a riflessioni sugli esempi (anche negativi) che si riscontrano nella letteratura. Soprattutto, abbiamo cercato di indicare gli aspetti legati al 'saper fare' statistica; sia con gli strumenti on line in rete, sia con il linguaggio di programmazione R, che si scarica gratuitamente dall'indirizzo `https://www.r-project.org/`. Da alcuni anni sta anche avendo rapida diffusione l'ambiente di sviluppo integrato (IDE) **R Studio**, `https://www.rstudio.com/`. Per esercitarci, ci siamo serviti dell'ambiente online R Fiddle, `http://www.r-fiddle.org/`.

Figura 1: R Fiddle, un ambiente online di programmazione in linguaggio R.

2 Perché ad un medico occorre uno statistico?

> *Partiamo da un presupposto: la statistica (quantomeno una comprensione di base della stessa) è fondamentale anche solo per comprendere l'attendibilità di quello che leggiamo e noi, secondo me, non possiamo davvero farne a meno; più che utile ci è necessaria.*

Questa è l'impressione di uno studente che ha seguito il corso. Io credo che la statistica in uno studio scientifico sia come il basso in un complesso musicale pop o jazz: se non c'è, se ne sente la mancanza. Certo, condivido pienamente l'opinione che 'non tutto è numero' e che la medicina deve essere un giusto equilibrio tra una **scienza evidence-based** e un'**arte osservativa** (come ci dice il professor Renzo Carretta). Certo che se il medico Benjamin Spock, autore di vari libri di pediatria negli anni 50, fosse stato più diligente e sistematico nella sua arte osservativa su come adagiare nel lettino i neonati, proni o bocconi, decine di migliaia di bimbi non sarebbero morti di SID [7].

Nella lezione-seminario sugli studi longitudinale abbiamo ricordato che il presidente della American Statistical Association, Ron Wasserstein, ha recentemente pubblicato **The ASA's statement on p-values: context, process, and purpose**, che si scarica gratuitamente dalla rete [11]. Vi raccomando assolutamente la lettura del paper. Inoltre, un articolo dal taglio 'operativo' che sempre raccomando di tenere a portata di mano, è quello di Douglas Curran-Everett e Dale Benos, **Guidelines for reporting statistics in journals published by the American Physiological Society** [5]. Da qui, si legge che:

- consultate uno statistico quando pianificate lo studio (= la vostra tesi di laurea).

3 Quale modello statistico adottare?

> *Ci potrebbe fornire un riassunto finale con le cose salienti da sapere, una griglia con una definizione e un esempio, per avere sempre uno specchietto da consultare nelle evenienze?*

È una domanda molto appropriata. Lo è a tal punto che Michael Crawley, autore di uno dei libri che amo, **Statistics: an introduction using R**, [4], lo fa alla pagina numero uno del libro! E ve lo dimostro nelle due pagine che seguono:

1

Fundamentals

The hardest part of any statistical work is getting started. And one of the hardest things about getting started is choosing the right kind of statistical analysis. The choice depends on the nature of your data and on the particular question you are trying to answer. The truth is that there is no substitute for experience: the way to know what to do is to have done it properly lots of times before.

The key is to understand what kind of *response* variable you have got, and to know the nature of your *explanatory* variables. The response variable is the thing you are working on: it is the variable whose variation you are attempting to understand. This is the variable that goes on the y axis of the graph (the ordinate). The explanatory variable goes on the x axis of the graph (the abscissa); you are interested in the extent to which variation in the response variable is associated with variation in the explanatory variable. A continuous measurement is a variable like height or weight that can take any real numbered value. A categorical variable is a *factor* with two or more *levels*: sex is a factor with two levels (male and female), and rainbow might be a factor with seven levels (red, orange, yellow, green, blue, indigo, violet).

It is essential, therefore, that you know:

- which of your variables is the response variable?
- which are the explanatory variables?
- are the explanatory variables continuous or categorical, or a mixture of both?
- what kind of response variable have you got – is it a continuous measurement, a count, a proportion, a time-at-death, or a category?

These simple keys will then lead you to the appropriate statistical method:

1. The explanatory variables (pick one of the rows):

(a) All explanatory variables continuous	*Regression*
(b) All explanatory variables categorical	*Analysis of variance (ANOVA)*
(c) Some explanatory variables continuous some categorical	*Analysis of covariance (ANCOVA)*

Statistics: An Introduction Using R, Second Edition. Michael J. Crawley.

2. The response variable (pick one of the rows):

(a) Continuous	*Regression, ANOVA or ANCOVA*
(b) Proportion	*Logistic regression*
(c) Count	*Log linear models*
(d) Binary	*Binary logistic analysis*
(e) Time at death	*Survival analysis*

There is a small core of key ideas that need to be understood from the outset. We cover these here before getting into any detail about different kinds of statistical model.

Alla luce di quello che ci insegna Crawley, proviamo a rifare questo esercizio. Abbiamo due gruppi di pazienti, bianco e nero, cui abbiamo misurato dei livelli di insulina, che vogliamo associare ai livelli misurati di retinolo:

	gruppo	insulina	retinolo
1	bianco	23	19
2	**nero**	**16**	**37**
3	bianco	34	23
..	..	..	..

- quale delle tre variabili è la variabile risposta?
- quali sono le variabili esplicative ('covariate')?
- le variabili esplicative ('covariate') sono continue, o categoriche, o una mistura di entrambe?
- quale tipo di risposta abbiamo - una misura continua, una 'conta', una proporzione, un tempo-alla-morte, o una categoria?

Sulla base di ciò, quale modello statistico è adeguato per studiare questi dati?

ANCOVA; retinolo, risposta continua.

4 Il t-test / la Anova

4.1 A cosa serve

	gruppo	retinolo
1	bianco	19
2	**nero**	**37**
3	bianco	23
..	..	..

Serve a 'trovare' se vi siano differenze, in media, tra vari gruppi. Più precisamente, se la risposta è una misura continua, e se l'unica variabile esplicativa è un fattore a due livelli (i.e. due gruppi di pazienti), allora il **test t di Student** potrebbe proprio fare al caso vostro. Se il fattore è a tre o più livelli (i.e. tre o più gruppi di pazienti), potreste ricorrere ad una **Anova**. Entrambi i test ricadono nell'ampia classe dei **modelli lineari**.

4.2 Come si fa con R

```
modello1 = lm(retinolo ~ 1 + gruppo)
summary(modello1)
confint(modello1, level = 0.90)
confint(modello1, level = 0.95)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.80	2.11	10.35	0.00
grupponero	6.40	2.98	2.15	0.06

	5 %	95 %
(Intercept)	17.88	25.72
grupponero	0.86	11.94
	2.5 %	97.5 %
(Intercept)	16.94	26.66
grupponero	-0.47	13.27

4.3 Quali sono gli errori da evitare

Ci ricordiamo che il t test 'funziona' se i dati sono distribuiti normalmente e se essi sono omoschedastici (i.e. egualmente 'sparpagliati'). La cosa migliore da fare è quella di eseguire la diagnostica del modello, con il comando:

```
par(mfrow = c(2,2))
plot(modello1)
```

Nel primo grafico, Residual vs. Fitted, osserveremo due 'gruppi' di punti, corrispondenti ai due livelli (nel nostro esempio, bianco e nero). Quello che si vuole è che la riga rossa sia allineata allo 0 (media nulla dei residui) e che sia pressoché orizzontale (assenza di *drift* dei dati). Si vuole inoltre che i due 'gruppi' siano approssimativamente della medesima ampiezza: buon segno di omoschedasticità. Il secondo pannello, Normal Q-Q, ci dice se c'è normalità dei residui: i pallini si devono grossomodo adagiare sulla diagonale tratteggiata, e non formare una 'gondola' o un 'serpente'. Il terzo pannello non ci interessa. Nel quarto pannello Residuals vs Leverage vengono messi in evidenza eventuali punti influenti, o outlier, del modello. Se ve ne fossero, la curva rossa non sarebbe orizzontale, e vi sarebbe qualche punto molto distante dagli altri esternamente ad una curva di tipo iperbolico: cattivo segno. Se nel plot diagnostico qualcosa non torna, contattate lo statistico di turno.

4.4 Cosa si scrive nella tesi

Per verificare se vi siano differenze tra i livelli medi di retinolo all'interno dei due gruppi abbiamo utilizzato un modello lineare (test t di Student). È risultato che il gruppo dei pazienti bianchi esprime una risposta media pari a 21.8 (s.e. 2.1) unità (intervallo di fiducia al 90 per cento compreso tra 17.9 e 25.7), mentre con fiducia del 90 per cento possiamo affermare che il gruppo dei pazienti neri differisce dai bianchi ($p = 0.06$) con una risposta media di 6.4 (s.e. 3.0) unità (da 0.9 a 11.9 con fiducia del 90 per cento).

4.5 E se i gruppi invece di due sono tre o più?

Semplice: si va su Google, si digitano le parole chiave massimo borelli anova confronti multipli e ci si legge la dispensa sull'Anova.

5 La retta di regressione

5.1 A cosa serve

	insulina	retinolo
1	23	19
2	16	37
3	34	23
..	..	..

Beh, banale, direi: a trovare la migliore ('migliore' nel senso del Teorema di Gauss e Markov) retta che attraversa una nuvola di punti. Molti pensano che i **modelli lineari** si chiamino così proprio perché c'è di mezzo la retta di regressione (costoro cambiano immediatamente idea quando vedono che **un** modello lineare potrebbe essere, ad esempio, **tre** parabole). Più precisamente, se la risposta è una misura continua, e se la variabile esplicativa è anche una misura continua, allora questo paragrafo fa per voi.

5.2 Come si fa con R

```
modello2 = lm(retinolo ~ 1 + insulina)
summary(modello2)
confint(modello2, level = 0.90)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.1657	6.9594	4.62	0.0017
insulina	-0.3213	0.3020	-1.06	0.3184

Residual standard error: 5.537 on 8 degrees of freedom, Multiple R-squared: 0.1239, Adjusted R-squared: 0.01444, F-statistic: 1.132 on 1 and 8 DF, p-value: 0.3184

	5 %	95 %
(Intercept)	19.22	45.11
insulina	-0.88	0.24

5.3 Quali sono gli errori da evitare

Nella regressione i residui ('errori') devono essere distribuiti normalmente, con media nulla (i.e. senza errori sistematici e senza drift) e deviazione standard costante (i.e. non eteroschedasticità). I guai provengono anche da eventuali punti isolati che abbiano rilevante forza di leva. La cosa migliore da fare è quella di eseguire la diagnostica del modello, con il comando:

```
par(mfrow = c(2,2))
plot(modello2)
```

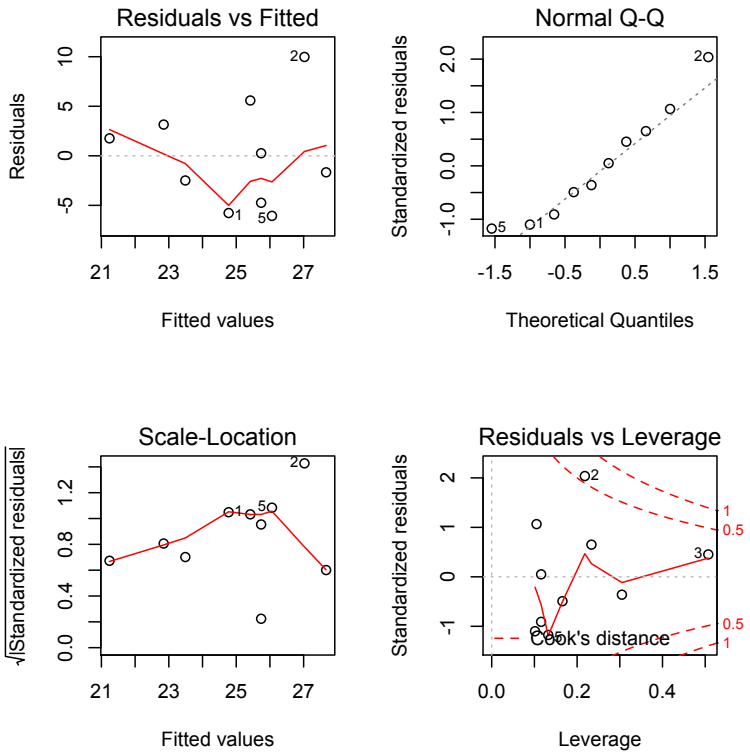


Figura 2: il plot diagnostico del modello statistico numero 2.

Nel primo grafico, Residual vs. Fitted, la riga rossa appare perturbata al centro. Sarà sicuramente colpa della piccola dimensione campionaria. Il secondo pannello, Normal Q-Q, è addirittura perfetto: i pallini si adagiano sulla diagonale tratteggiata e non formano

alcuna 'gondola' o 'serpente'; i residui quindi si distribuiscono normalmente, come da sempre desideriamo. Il quarto pannello Residuals vs Leverage mostra che il punto numero 2 è un pochino 'strano', ma non ha un'eccezionale forza di leva. Conclusione: il modello 2 è proprio quello giusto da utilizzare nella vostra tesi.

5.4 Cosa si scrive nella tesi

Per verificare se i livelli medi di retinolo siano correlati ai livelli medi di insulina, abbiamo utilizzato un modello lineare (retta di regressione):

$$retinolo = \alpha + \beta \cdot insulina + \varepsilon$$

Dal campione preso in esame è risultato che le variazioni di insulina ($\beta = -0.32, s.e. = 0.30$) non modificano in maniera significativa (p value = 0.32) i livelli medi di retinolo (errore standard residuale $\varepsilon \sim N(0, \sigma), \sigma = 5.5$ su 8 gradi di libertà).

6 La Ancova

6.1 A cosa serve

Lo abbiamo ricordato qualche pagina fa nell'esempio introduttivo: se la risposta è una variabile continua, e le variabili esplicative sono una mistura di variabili categoriche e continue, allora l'Analisi della Covarianza è la tecnica che vi può servire.

Ma in un modello statistico, bisogna fare molta attenzione tra la parola **variabile esplicativa**, sinonimo di **covariata**, e la parola **predittore**. Nel vostro dataset, tutte le colonne di dati (eccetto la risposta, ovviamente) sono delle covariate. Ma non è affatto detto che tutte quelle covariate siano dei predittori della risposta: potrebbe infatti accadere (e succede quasi sempre, direi) che qualche covariata sia altamente correlata alle altre, risultando ridondante nel modello.

Facciamo un esempio. Nel paper [8] vengono riportati alcuni modelli statistici. La domanda è: a quale scopo si introducono nei modelli statistici covariate (il valore in scala logaritmica dell'antigene carcinoembrionario, CEA) che non siano predittori, contravvenendo al principio di parsimonia di Occam? La mia risposta è: non lo so, e non condivido questa scelta.

Table 4
Multivariate logistic regression model showing association of biomarkers with malignancy.

	OR	95% CI		p
Premenopausal				
HE4 (log)	2.13	0.87	5.20	0.098
CA125 (log)	1.27	0.81	2.00	0.292
CEA (log)	1.44	0.72	2.86	0.300
Age at diagnosis	0.96	0.87	1.05	0.391
Postmenopausal				
HE4 (log)	4.17	1.36	12.77	0.012
CA125 (log)	1.43	0.89	2.28	0.136
CEA (log)	0.50	0.21	1.19	0.117
Age at diagnosis	0.89	0.82	0.96	0.004
Combined (pre- and postmenopausal)				
HE4 (log)	2.60	1.34	5.04	0.005
CA125 (log)	1.30	0.95	1.78	0.096
CEA (log)	0.93	0.57	1.52	0.779
Age at diagnosis	0.99	0.96	1.03	0.708

OR = odds ratio, CI = confidence interval.

Figura 3: covariate o predittori?

Sarebbe forse utile provare a fare qualche 'simulazione' pratica ..

6.2 Come si fa con R: dal modello massimale al modello minimale adeguato

Dovremmo forse, in primis, indugiare in sottili distinzioni tra Ancova e Anova, two-way e one-way, with interaction e without interaction.. ma noi siamo pragmatici e non lo facciamo! L'unica cosa da fare, prima di proseguire, è leggere su Wikipedia la voce Rasoio di Occam (la versione in inglese è molto ben fatta). Compresa che sia la necessità di non introdurre in un modello delle covariate che siano ridondanti, proviamo ad esercitarci a trovare il **modello minimale adeguato** che descriva il mio dataset `diabete.csv`:

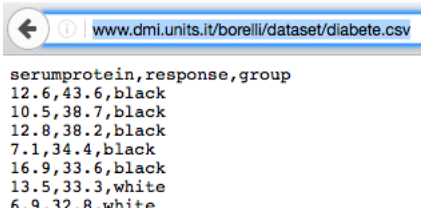
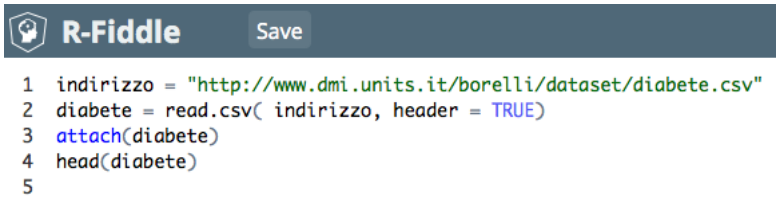


Figura 4: il dataset `diabete.csv` on line sul mio sito web

Proviamo ad esercitarci con R-Fiddle (vedi Figura 1), importando direttamente dalla rete il dataset, e visualizzandone le prime sei righe, con questi comandi:

```
indirizzo = "http://www.dmi.units.it/borelli/dataset/diabete.csv"
diabete = read.csv( indirizzo, header = TRUE)
attach(diabete)
head(diabete)
```



Se abbiamo digitato tutto correttamente e se schiacciamo il pulsante verde Run Code, otteniamo in basso in colore arancione le prime sei righe del dataset:

Graphs Run Code

```
  serumprotein response group
1         12.6     43.6 black
2         10.5     38.7 black
3         12.8     38.2 black
4          7.1     34.4 black
5         16.9     33.6 black
6         13.5     33.3 white
>
```

Ora, partiamo con il **modelloA**, un **modello massimale**: tutte le variabili esplicative (che sono poi solo due, `serumprotein` e `group`) vengono considerate potenziali predittori della risposta `response`. Questo significa che la `response` dipende dalla `serumprotein`, come nel paragrafo **5 La retta di regressione**; ma siccome vi sono due `group`, avremo dunque due rette di regressione, e vorremo capire - come nel paragrafo **4 il t-test**, se questi due gruppi si comportano in modo diverso o no. Se i due gruppi si comporteranno in modo diverso, allora vuol dire che appartenere ad uno o all'altro `group` fornisce due diverse informazioni alla `serumprotein`. In termini geometrici, questo comporterà che le due rette di regressione saranno diverse, sia come intercetta che come pendenza.

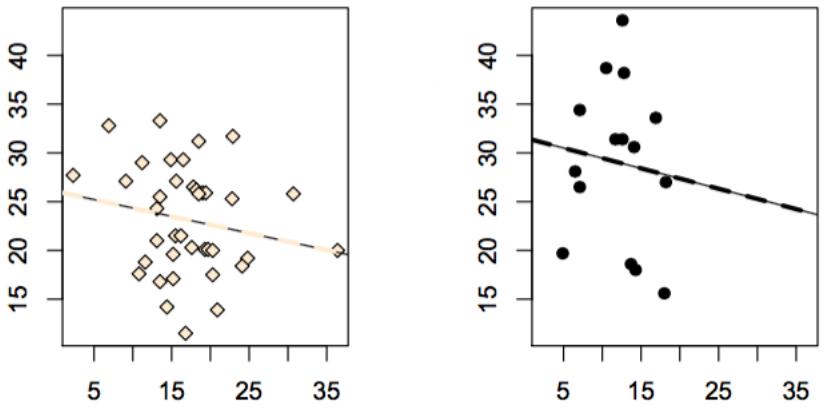
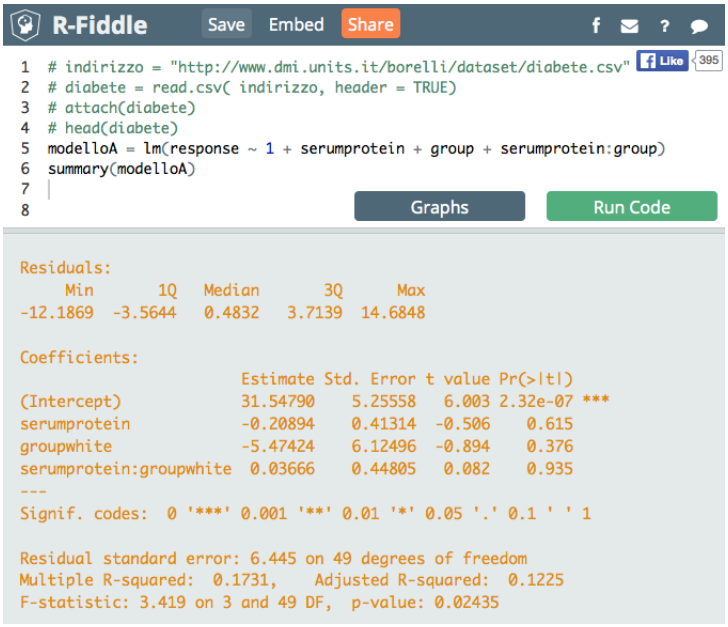


Figura 5: il modello massimale: due rette di regressione, entrambe con differenti pendenze e differenti intercette.

Indicheremo con il simbolo `serumprotein:group` questa influenza tra il `group` e la `serumprotein`. Sono d'accordo con voi che la sintassi sia oscura: si chiama notazione di Wilkinson e Rogers (ma è solo una della millanta cose oscure che ci sono in Statistica):

```
modelloA = lm(response ~ 1 + serumprotein + group + serumprotein:group)
summary(modelloA)
```

Attenzione! Per non fare pasticci, modifichiamo le prime quattro righe di comando inserendoci dei diesis / cancelletti, che rappresentano il carrattere di commento in R: questo farà sì che di volta in volta non verrà ri-caricato il dataset (vedrete che i comandi diventano di colore verde). Poi, copiamo quelle due righe di comando e schiacciamo il pulsante verde Run Code.



Leggiamo le `Estimate` ed interpretiamo quei coefficienti, tenendo sott'occhio la Figura 5. La retta dei 'bianchi' ha equazione:

$$response = 31.55 - 0.21 \cdot serumprotein$$

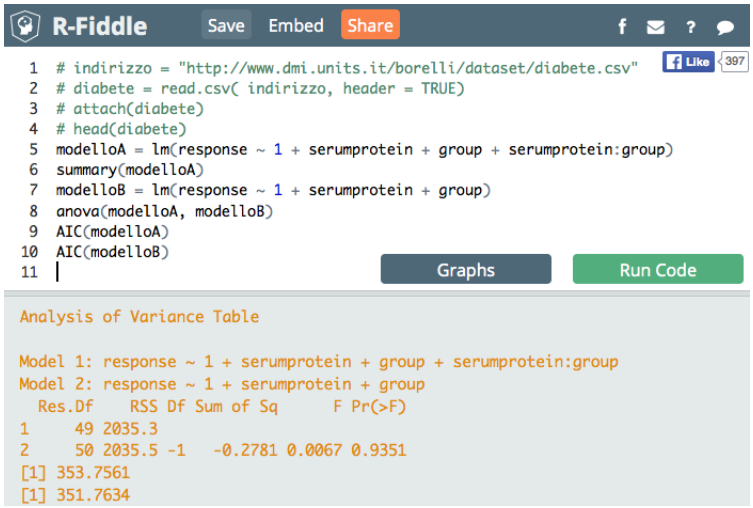
mentre la retta dei neri ha equazione:

$$response = 26.07 - 0.17 \cdot serumprotein$$

La domanda sorge spontanea: siccome - anche la Figura 5, ad occhio, ce lo fa intuire - la pendenza 0.21 della retta dei bianchi e la pendenza 0.17 della retta dei neri sono praticamente lo stesso numero, stai a vedere che in realtà esse sono parallele? Un forte indizio ce lo da anche il p-value 0.935 del coefficiente di interazione `serumprotein:group`.

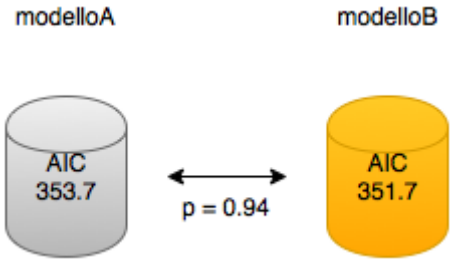
Per fare questo controllo, impostiamo un `modelloB` additivo e facciamo due verifiche: un'**analisi della devianza** con il comando `anova` e un'analisi dei **criteri di informazione** con il comando `AIC`:

```
modelloB = lm(response ~ 1 + serumprotein + group)
anova(modelloA, modelloB)
AIC(modelloA)
AIC(modelloB)
```



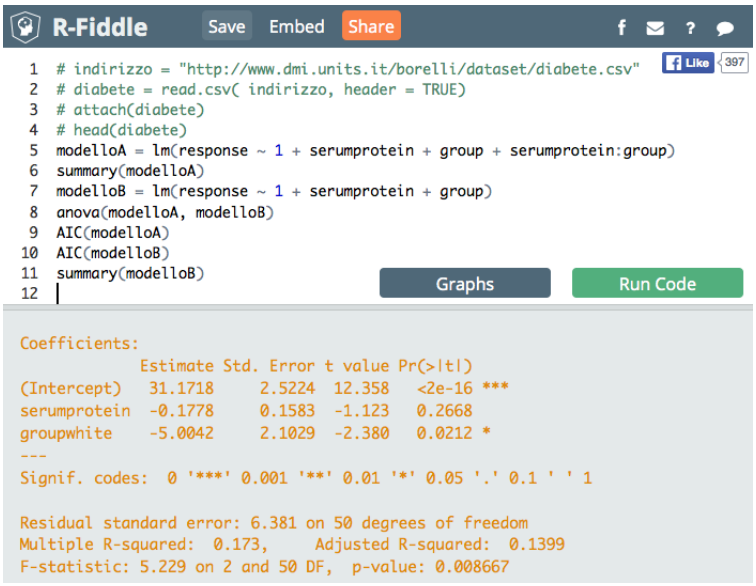
Leggete su Wikipedia chi è stato il professot Hirotsugu Akaike, e cosa significa il suo criterio di informazione, che funziona all'insegna di *'small is beautiful'*, 'nelle botti piccole ci sta il vino buono'. Siccome il `modelloB` ha un criterio di informazione inferiore a quello del `modelloA`, e siccome i due modelli non sono differenti in senso statistico tra di loro (p-value 0.9351), allora è meglio preferire il `modelloB`, che vi lascia un grado di libertà in più (DF 50 contro DF 49), e quindi vi costa un parametro di meno:

modello	effetti fissi	effetti casuali
modelloA	4 (due pendenze, due intercette)	1 (residual standard error ε)
modelloB	3 (stessa pendenza, due intercette)	1 (residual standard error ε)



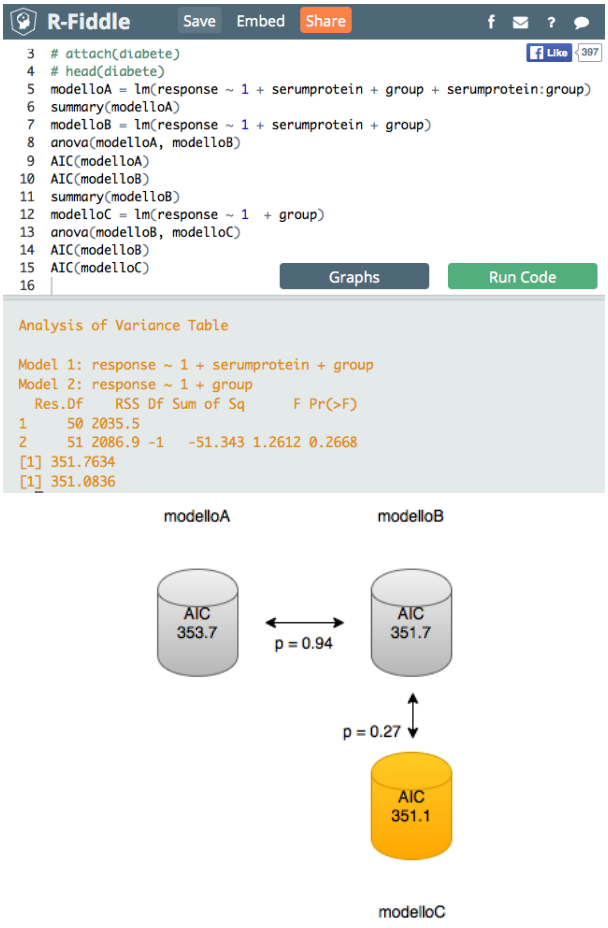
Esaminiamo tuttavia il `summary` del `modelloB`:

```
summary(modelloB)
```



È vero che inziano ad apparire delle stelline, e il modello 'è significativo' diremmo in maniera grossolana (p-value = 0.009). Ma, attenzione, il termine `serumprotein`, che è la pendenza delle due rette di regressione, non è significativa. Significa forse che noi dovremmo levar via quella pendenza? Proviamo:

```
modelloC = lm(response ~ 1  + group)
anova(modelloB, modelloC)
AIC(modelloB)
AIC(modelloC)
```



Ebbene sì, quella pendenza è ridondante, e le due rette di regressione sono in realtà orizzontali; se chiedete il `summary` del `modelloC`, vedrete che per i neri:

$$response = 29.0$$

mentre per i bianchi:

$$response = 23.1$$

Se riguardate il capitolo **4 Il t-test** vedrete che questi due numeri non sono altro che i livelli medi di retinolo nei due gruppi di pazienti, che stavolta differiscono in maniera altamente significativa:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.0267	1.6517	17.57	0.0000
groupwhite	-5.9004	1.9506	-3.02	0.0039

modello	effetti fissi	effetti casuali
modelloA	4 (due pendenze, due intercette)	1 (residual standard error ε)
modelloB	3 (stessa pendenza, due intercette)	1 (residual standard error ε)
modelloC	**2** (nessuna pendenza, due intercette)	**1** (residual standard error ε)

6.3 Cosa si scrive nella tesi

L'analisi dei dati sul dataset `diabete` ci mostra che la `response` può venir associata al `group` ma non alla `serumprotein`. Avendo ipotizzato un modello lineare massimale e, mediante una procedura di tipo top-down, avendo selezionato il modello minimale adeguato in termini di criteri di informazione e di significatività dei termini della regressione, ci risulta che il `group white` abbia una risposta media inferiore di circa 5.9 (s.e. 2.0, p = 0.004) unità rispetto al `group black`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.03	1.65	17.57	< 0.001
groupwhite	-5.90	1.95	-3.02	0.004

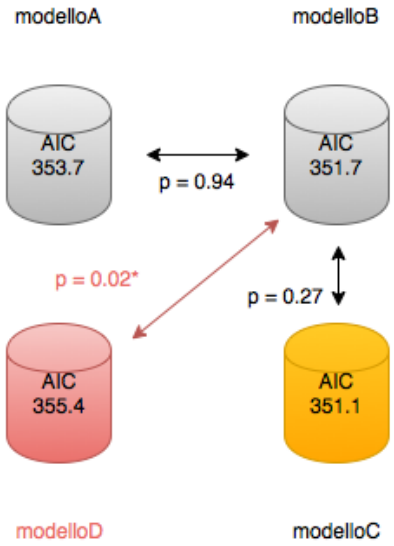
6.4 Quali sono gli errori da evitare

Non effettuare una selezione del modello accurata, non basandosi sui criteri di informazione o sull'analisi della devianza, vi può condurre a prendere cantonate colossali. Infatti il `modelloD`, che 'è significativo':

```
modelloD = lm(response ~ 1  + serumprotein)
summary(modelloD)
anova(modelloB, modelloD)
AIC(modelloD)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.8255	2.5681	11.61	0.0000***
serumprotein	-0.3207	0.1530	-2.10	0.0411*

si rivela essere un pessimo modello statistico, che non interpreta correttamente i dati: differisce significativamente dal `modelloB`, ha un criterio di informazione di Akaike superiore a tutti i modelli, ed ha un errore standard dei residui molto elevato, $\sigma_\varepsilon = 6.66$.

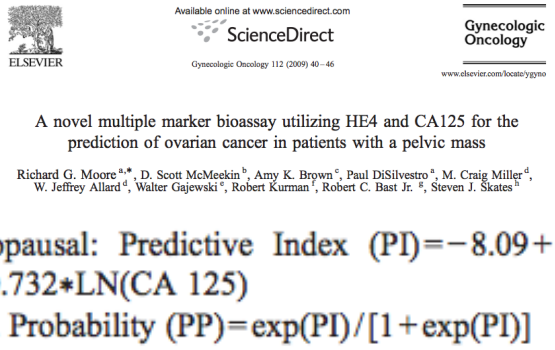


E perciò a nessuno, e dico nessuno, dovrebbe mai venire in mente di piazzare il `modelloD` in una tesi di laurea. Men che meno, in un paper. E ho detto tutto!

7 La regressione logistica

7.1 A cosa serve

Quando la risposta è una variabile categorica di tipo binomiale, e le variabili esplicative sono una mistura di variabili categoriche e continue, allora la regressione logistica è la tecnica che potrebbe fare al caso vostro.
Ad esempio, l'indice ROMA (Risk of Ovarian Malignancy Algorithm) utilizza l'espressione di due proteine per predire la malignità (variabile categorica di tipo binomiale) del tumore ovarico:



ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect
Gynecologic Oncology 112 (2009) 40–46

Gynecologic Oncology
www.elsevier.com/locate/ygyno

A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass

Richard G. Moore [a,*], D. Scott McMeekin [b], Amy K. Brown [c], Paul DiSilvestro [a], M. Craig Miller [d], W. Jeffrey Allard [d], Walter Gajewski [e], Robert Kurman [f], Robert C. Bast Jr. [g], Steven J. Skates [h]

Postmenopausal: Predictive Index (PI) = −8.09 + 1.04∗LN (HE4) + 0.732∗LN(CA 125)
Predicted Probability (PP) = exp(PI)/[1 + exp(PI)]

Come si vede, la formula fornisce un *indice predittivo* del tipo $PI = a + bx_1 + cx_2$, che è in definitiva un modello lineare, se pensiamo ad x_1 e x_2 come i logaritmi dei marker HE4 e CA125. Ma la formula matematica 'non finisce qui', perchè dall'indice predittivo PI si passa alla probabilità di esito maligno della neoplasia (Predicted Probability) per mezzo di una **funzione di collegamento** (funzione di **link**):

$$p = \frac{e^{PI}}{e^{PI} + 1}$$

Non si tratta di una roba matematica che cade dal cielo, ma si tratta della formula inversa del cosiddetto **logit**, che si definisce come $\log(\frac{p}{1-p})$, che ha un caratteristico andamento di tipo sigmoidale e che assume valori sempre compresi tra 0 ed 1, come accade per ogni misura di probabilità (guardate la Figura 6 alla prossima pagina).
Dunque, nella **regressione binomiale**, ci servirà determinare:

- una 'formula' che coinvolga in maniera lineare i predittori, come abbiamo visto nel capitolo precedente **6 La Ancova**

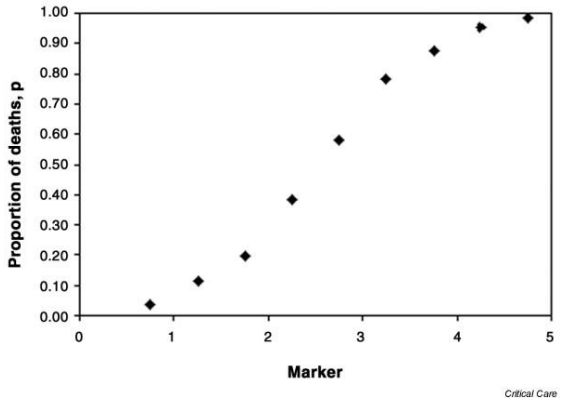


Figura 6: Logistic regression, http://ccforum.com/content/9/1/112

- una 'funzione di collegamento' che trasformi l'output lineare (gaussiano) in una distribuzione binomiale (e nel nostro caso, per default, sarà `link = 'logit'`)
- una 'famiglia di variabili aleatorie' binomiali per modellare i residui (e nel nostro caso sarà `family = 'binomial'`)

7.2 Un esempio di regressione logistica

Come esempio, scegliamo il dataset `ovarian` raccolto dal collega professor Ricci e dalla dottoressa Shadi Nagaf, che coinvolge 210 donne:

```
www <- "http://www.dmi.units.it/borelli/dataset/ovarian.csv"
ovarian = read.csv( www , header = TRUE )
attach(ovarian)
```

Il dataset raccoglie 4 marker proteici espressi in diversi tipi di neoplasie del sistema genitale femminile, l'età delle pazienti, la loro condizione di fertitilità e la diagnosi istologica al prelievo bioptico delle masse reperite all'indagine ecografica. Ci proponiamo di scoprire quale sia il ruolo predittivo delle covariate del dataset e se questo sia in accordo con la letteratura (indice ROMA).
Per prima cosa trasformiamo i dati relativi ai marker in scala logaritmica per ridurre la considerevole asimmetria dei loro valori, ed ipotizziamo un modello massimale:

```
logHE4 = log(HE4/100)
logCA125 = log(CA125/100)
```

```
logCA199 = log(CA199/100)
logCEA = log(CEA/100)
modelloMassimale = glm(OUTCOME ~ 1 + logHE4 + logCA125 + logCA199 + logCEA
+ ETA + MENOPAUSA, family = binomial)
summary(modelloMassimale)
```

Adesso, per esercizio, provate a fare da voi la selezione del modello, come abbiamo imparato nel capitolo precedente. Per selezionare il modello minimale adeguato servitevi certamente dei criteri di informazione di Akaike:

`AIC(` .. *modello* .. `)`

ma per selezionare due modelli tra di loro in base all'analisi della devianza, tenete presente che dovete utilizzare la variabile aleatoria del Chi quadrato, con questa sintassi:

```
anova (modelloGrande, modelloPiccolo, test = "Chisq")
```

Se eseguite tutto correttamente, dovreste giungere al seguente modello minimale:

```
modelloMinimale = glm(OUTCOME ~ logHE4 + logCA125, family = binomial)
summary(modelloMinimale)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-15.4469	2.7189	-5.68	0.0000
logHE4	2.7759	0.6300	4.41	0.0000
logCA125	0.6352	0.1979	3.21	0.0013

La cosa interessante appare dal raffronto con il modello della letteratura:

```
modelloLetteratura = glm(OUTCOME ~ logHE4 + logCA125 + MENOPAUSA,
family = binomial)
summary(modelloLetteratura)
```

Cosa c'è di interessante? C'è che nel nostro campione, diversamente da quanto affermato dagli studi esistenti, la condizione di menopausa non appare essere un predittore significativo (p value = 0.0998. Ricordiamoci che con un campione di più di 200 donne non è realistico assumere un livello α del 10 per cento) e perciò, in base al principio di Occam, sarebbe opportuno non considerarla. Ed infatti, il `modelloLetteratura` ed il `modelloMinimale` non differiscono in senso significativo:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.4210	2.8131	-4.77	0.0000
logHE4	2.3336	0.6527	3.58	0.0003
logCA125	0.6848	0.2029	3.37	0.0007
MENOPAUSAPRE	-0.9380	0.5699	-1.65	0.0998

```
anova(modelloLetteratura, modelloMinimale, test = "Chisq")
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	206	107.33			
2	207	110.11	-1	-2.78	0.0955

Tuttavia, i criteri di informazione suggeriscono che il `modelloLetteratura` sia preferibile

```
AIC(modelloLetteratura)   # 115.3
AIC(modelloMinimale)      # 116.1
```

7.3 Cosa si scrive nella tesi

L'analisi dei dati sul nostro campione conferma i risultati della letteratura: il rischio di neoplasia ovarica è predetto dai marker HE4 ($p < 0.001$) e CA125 ($p = 0.001$). Al contrario, gli altri biomarcatori considerati non appaiono associati alla patologia. In particolare, non è chiaro il ruolo della condizione di menopausa, che non appare essere un predittore (p value = 0.10) ma da un punto di vista informativo esibisce un ruolo di rilievo (AIC 115.3 versus AIC 116.1).

7.4 Quali sono gli errori da evitare

Ci dobbiamo ricordare che con la variabile aleatoria binomiale, che abbiamo utlizzato in questa regressione, lo sperimentatore non può scegliere ad arbitrio la media e la deviazione standard (come invece accade nella gaussiana). In questo caso media e deviazione standard sono invece predeterminate dalla numerosità campionaria e dalla probabilità dell'evento considerato. Questo comporta [6] che bisogna prestare particolare attenzione all'output del comando `summary`:

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.58  on 209  degrees of freedom
```

```
Residual deviance: 110.11  on 207  degrees of freedom
```

Quello che interessa, brevemente, è che la devianza residua sia inferiore ai gradi di libertà. In questo caso la devianza residua vale 110.11, e questo numero è inferiore ai gradi di libertà che sono 207 (infatti, il campione raccoglieva 210 donne e il modello ha 3 parametri). Se così non fosse e se avessimo una devianza residua superiore ai gradi di libertà, dovremmo modificare il parametro di dispersione della famiglia binomiale utilizzando il comando

```
family = "quasibinomial"
```

invece che `family = binomial`. Il software R provvede in maniera autonoma a fissare il parametro opportuno per mezzo di una procedura iterativa. Si noterà nell'output che le stime e gli standard error non vengono modificati, ma cambiano i p-value dei coefficienti, che di norma diventano 'meno generosi', 'meno significativi'.

8 Analisi di sopravvivenza

8.1 A cosa serve

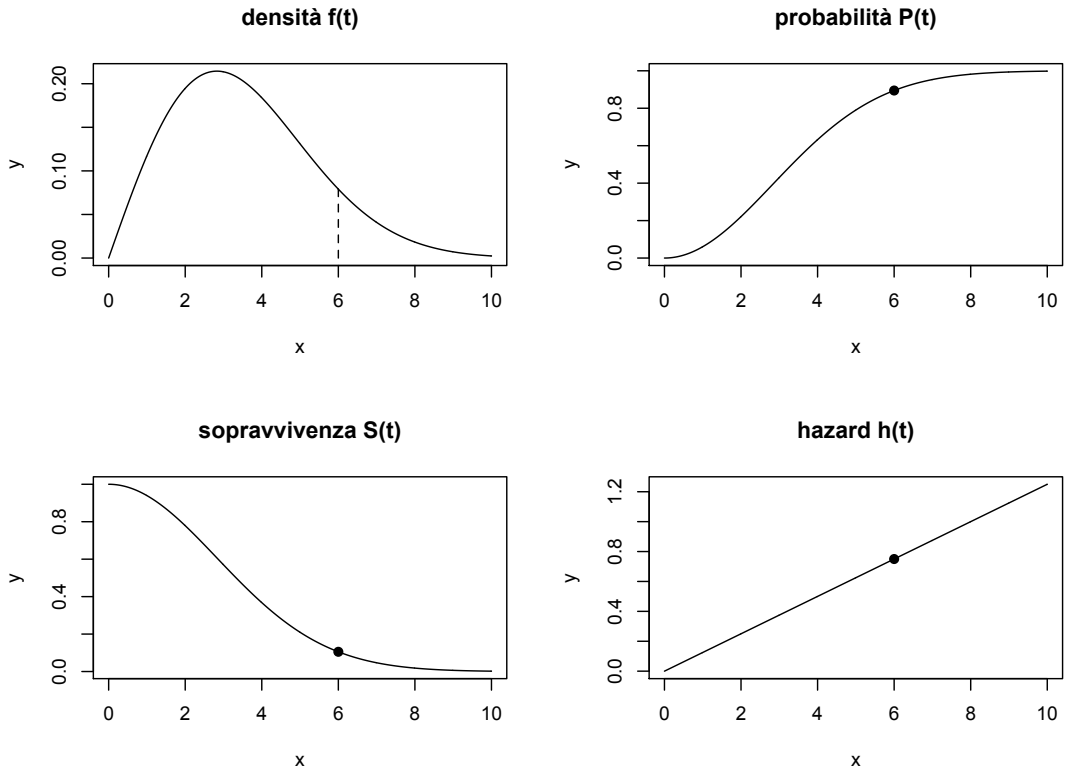
Consideriamo il dataset `tiroide`, gentilmente messo a disposizione da Marina e da Simone, in cui sono riportati dati clinici relativi a 57 pazienti che hanno subito un trattamento chirurgico alla tiroide, con o senza interessamento ai linfonodi (`LCC`, `no` / `si`):

	Mesi	Statoattuale	LCC	Eta
1	108	0	no	71
2	39	0	no	28
3	94	0	no	57
4	23	1	no	58
5	78	0	no	62
..	..	..	..	..
57	10	0	si	37

L'analisi di sopravvivenza serve a stabilire se il periodo di follow-up, qui indicato con `Mesi`, si differenzi in associazione al diverso trattamento `LCC`, oppure ad altre covariate considerate nello studio, come ad esempio l'`Eta` dei pazienti. È importante notare, dal punto di vista matematico, che l'evento di interesse è la condizione `1`, deceduto, nel fattore `Statoattuale`. L'evento complementare, `0`, in realtà potrebbe voler indicare due fatti: o che il paziente, ad oggi, è ancora vivo; oppure, che l'ultima volta che è stato visitato risultava -ovviamente- vivo, ma da allora non abbiamo più notizie di lui (i cosiddetti, **dati censurati a destra**).

8.2 Cosa c'è da sapere

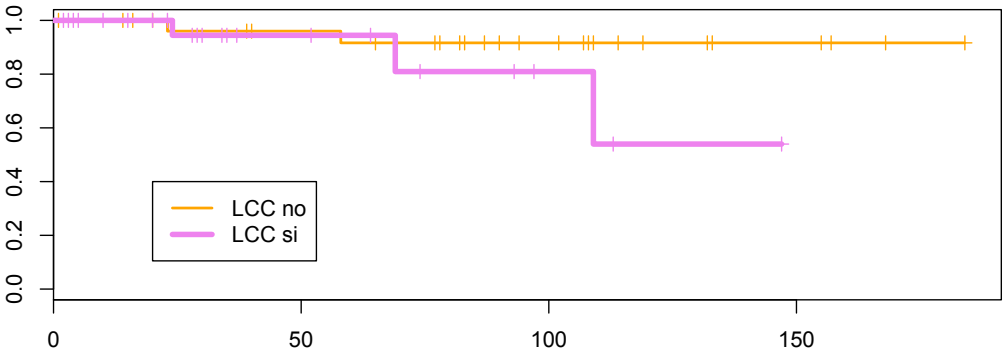
C'è da sapere che, a differenza dei modelli lineari, nei quali abbiamo a che fare per esempio con la variabile aleatoria normale, la sua densità, e la probabilità individuata da un certo quantile, qui invece avremo a che fare con la `sopravvivenza` e con l'`hazard` (i.e. il rischio). Vediamo con una figura come sono legati tra loro questi concetti.



Mettiamo il caso che abbiamo seguito un gran numero di pazienti, e abbiamo visto che l'outcome sfavorevole sopraggiunge molto frequentemente circa dopo tre anni l'insorgenza della patologia, e che dopo circa in dieci anni nessuno sopravvive. La curva di **densità** $f(t)$ non è altro che la rappresentazione matematica dell'istogramma degli eventi.
Se ci chiediamo quale possa essere la percentuale di persone che hanno avuto outcome sfavorevole nei primi 6 anni (integrale = area a sinistra di 6 nella densità), lo possiamo leggere dal grafico della **probabilità** $P(t)$: il pallino nero ci indica uno $0.90 = 90\%$ di persone.
Se invece ci chiediamo possa essere la percentuale di persone che sopravviveranno ai primi 6 anni, lo leggiamo dal grafico della **sopravvivenza** $S(t)$, che è esattamente il grafico della probabilità $P(t)$ capovolto; infatti, il pallino nero vale qui circa $0.10 = 1 - 0.90$.
Ma dal punto di vista dei software risulta più agevole valutare l'**hazard** $h(t)$, che si può ottenere calcolando il rapporto $f(t)/S(t)$. Se vedete, infatti, la barretta tratteggiata $f(6)$ vale circa 0.08, e $f(6)/S(6) = 0.08/0.10 = 0.8$. C'è anche da dire che qui l'hazard appare come una retta solo perchè abbiamo scelto come densità una particolare variabile aleatoria $f(t)$ (molto usata) che si chiama di Weibull.

8.3 Come si fa con R

```
kaplan = survfit(Surv(Mesi, Statoattuale) ~ LCC)
plot(kaplan, col = c("orange", "violet"), lwd = c(2,4))
legend(20, 0.4, legend = c("LCC no", "LCC si"),
        lwd = c(2,4), col = c("orange", "violet"))
```



Le due curve di Kaplan e Meier stimano la sopravvivenza $S(t)$ nei gruppi `no` (arancione sottile) e `si` (violetta grossa) di `LCC`. Abbiamo ora almeno due validi modelli statistici per decidere se la sopravvivenza nei due gruppi sia diversa, e in ragione di quali variabili esplicative considerate nel dataset.

8.3.1 Metodo semiparametrico di Cox, o dei rischi proporzionali

```
modello = coxph(Surv(Mesi, Statoattuale) ~ LCC + Eta)
summary(modello)
```

fattore	coef	exp(coef)	se(coef)	z	Pr(> |z|)	lower .95	upper .95
LCCsi	1.00	2.72	0.93	1.07	0.28	0.44	16.91
Eta	0.06	1.06	0.04	1.64	0.10	0.99	1.15

8.3.2 Cosa si scrive nella tesi

L'analisi effettuato con il modello regressivo di Cox ci indica che i due trattamenti `LCC` non appaiono modificare il rischio di sopravvivenza: il rapporto tra i rischi $h_{si}(t)/h_{no}(t)$ vale approssimativamente 2.7, ma non possiamo affermare con una fiducia del 95 per cento che questa differenza sia significativa (intervallo di fiducia al 95 per cento del rapporto $h_{si}(t)/h_{no}(t)$ 0.44 - 16.91). Al contrario, l'età dei pazienti appare avere un maggior ruolo: per ogni anno di età l'incremento di rischio è di circa il sei per cento ($\exp(\beta) = 1.06$, intervallo di fiducia al 95 per cento 0.99 - 1.15)

8.3.3 Metodo parametrico, o del fattore di accelerazione

```
modello = survreg(Surv(Mesi, Statoattuale) ~ LCC + Eta)
summary(modello)
```

In questo caso, è indispensabile tenersi uno statistico a fianco e condurre le analisi in sua compagnia.

8.4 Quali sono gli errori da evitare

Con il modello `survreg` del fattore di accelerazione non ci sarà alcun errore, perchè avrete fatto tutto assieme ad uno statistico esperto. Con il modello dei rischi proporzionali `coxph` è importante non confondere il concetto di hazard $h(t)$ con la sopravvivenza $S(t)$; se nella tesi avete scritto che per ogni anno di età l'incremento di rischio nel gruppo LCC sì in rapporto al gruppo LCC no è di circa il sei per cento, questo **non** vuol dire che la probabilità di avere un outcome sfavorevole aumenta del sei per cento ogni anno: la relazione tra $h(t)$ ed $S(t)$ non è banale:

$$S(t) = \exp\left(-\int_0^t h(s)ds\right)$$

Se volete farvi una cultura di base potete per esempio cominciare dal libro di Broström [2].

9 Dai design cross-section ai design con misure longitudinali

Attenzione! In tutte le pagine precedenti abbiamo dato per scontato che abbiamo raccolto dati sui nostri pazienti 'facendo una fotografia della situazione esistente'. Caratterizziamo ogni paziente mediante un unico evento risposta, al quale associamo un certo numero di covariate. Se però tale evento risposta viene seguito nel tempo e misurato più e più volte (**design longitudinale**), allora la musica cambia. E lo vediamo con un esempio sciocco ma illuminante:

9.1 Lo strano caso delle gemelle Alice ed Ellen

Le gemelle Alice ed Helen sono due anziane signore che, dopo aver condotto una vita artistica di grande successo, decidono di riprendere gli studi di biostatistica che avevano interrotto alcuni decenni fa. Alice ed Ellen decidono di fare uno **studio osservazionale**: alzarsi dal letto assieme ogni mattina e immediatamente pesarsi, per rispondere alla seguente domanda: *Alice ed Ellen hanno lo stesso peso?*

All'indomani, eseguito il primo esperimento e preso nota del responso della bilancia (accuratissima, digitale, che non si lascia perturbare dalle onde gravitazionali, ecc. ecc.) la situazione è la seguente:

Alice	Ellen
73.60	73.80

A questo punto, Alice ed Ellen sarebbero propense a decidere *che* ***non hanno*** *lo stesso peso*, giacché, ragionando da un punto di vista puramente matematico, i due numeri non coincidono.

Ma le gemelle sanno che, nella Natura, la variabilità la fa da padrona[1] e così scelgono di fare un secondo esperimento, ossia di pesarsi per cinque mattine consecutive (**studio osservazionale longitudinale**):

	Alice	Ellen
1	73.60	73.80
2	73.40	73.50
3	74.10	74.60
4	73.50	73.80
5	73.20	73.60

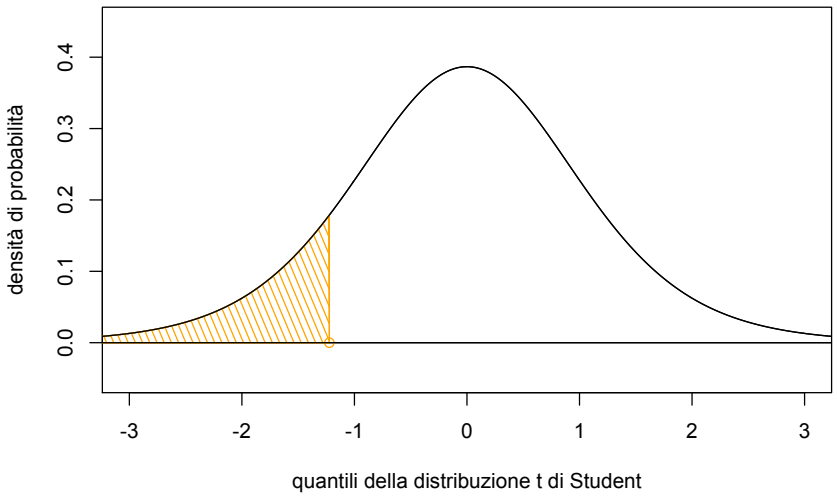
Per dirimere la questione esse ricorrono al celebre test t di Student. Come tutti ricordano, si vuole decidere se la media dei pesi di Alice sia diversa 'in senso statistico' dalla media dei pesi di Ellen, immaginando che per ciascuna di esse siano stati osservati cinque numeri casuali provenienti da due variabili aleatorie gaussiane, di media (nel senso di valore atteso, o speranza matematica) diversa ma con la medesima dispersione (nel senso di deviazione standard, ovvero della varianza).

Ecco qui di seguito il listato dei comandi per eseguire il test con R. Nella pagina seguente riportiamo invece l'output fornito dal software.

```
> alice = c(73.6, 73.4, 74.1, 73.5, 73.2)
> ellen = c(73.8, 73.5, 74.6, 73.8, 73.6)
> t.test(alice, ellen, var.equal = TRUE)
```

```
Two Sample t-test

data:  alice and ellen
t = -1.2227, df = 8, p-value = 0.2562
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.865794  0.265794
sample estimates:
mean of x mean of y
    73.56     73.86
```

Alice ed Ellen ora sarebbero propense a decidere *che* ***hanno*** *lo stesso peso*, in base al `p-value = 0.2562` (non inferiore al 5%). Infatti, la differenza tra il peso medio di Alice (`73.56`) e quello di Ellen (73.86) dà luogo ad un consuntivo `t = -1.2227`, il quale rispetto alla variabile aleatoria `t` di Student a `df = 8` gradi di libertà (5 pesi di Alice + 5 pesi di Ellen - 1 valor medio di Alice - 1 valor medio di Ellen), equivale ad un'area di probabilità p pari a 0.2562, come vediamo nella regione tratteggiata della figura sottostante.



Le gemelle tuttavia ricordano che l'affidabilità delle misure aumenta con il numero di repliche. Scelgono perciò di continuare a pesarsi complessivamente per tre settimane,

	Alice	Ellen		Alice	Ellen
1	73.60	73.80	12	74.10	74.60
2	73.40	73.50	13	73.60	73.80
3	74.10	74.60	14	73.40	73.60
4	73.50	73.80	15	74.10	74.40
5	73.20	73.60	16	73.50	73.70
6	74.00	74.40	17	73.20	73.50
7	73.60	73.80	18	74.00	74.40
8	73.30	73.50	19	73.60	73.90
9	74.20	74.30	20	73.30	73.60
10	73.60	73.90	21	74.20	74.50
11	73.40	73.60	-	-	-

dando luogo al loro terzo esperimento. Nella pagina che segue riportiamo la tabella con i dati grezzi dei pesi e il risultato del relativo test t di Student.

```
Two Sample t-test

data:  peso by gemella
t = -2.4594, df = 40, p-value = 0.01834
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.51183215 -0.05007261
sample estimates:
mean in group alice mean in group ellen
           73.66190            73.94286
```

Colpo di scena! Alice ed Ellen ora si sentono confuse più che mai, perché ora dovrebbero decidere *che* ***non hanno*** *lo stesso peso*, in base al `p-value = 0.01834`, significativo; contrariamente a quello che era accaduto nel secondo esperimento. Tutto ciò è molto strano. Di chi è la colpa?

9.2 Tutta colpa di Student

Alice ed Ellen hanno commesso un errore: non si sono ricordate che il test t è appropriato quando siamo in presenza di dati indipendenti e non, come in questo caso, di **dati correlati** [10], come è tipico nei design sperimentali di tipo longitudinale in cui si eseguono **misure ripetute**, in tempi successivi, sul medesimo soggetto. Il test t di Student

invece è un **modello lineare ad effetti fissi**. Questo significa che detto $\mu = 73.66$ il peso medio di Alice ottenuto nel terzo esperimento, il peso medio di Ellen è superiore a quello della gemella di una costante **(effetto fisso)** $\beta_2 = 0.28 = 73.94 - 73.66$ (mentre per quello di Alice possiamo per completezza porre l'effetto fisso $\beta_1 = 0$). E, di volta in volta, i pesi delle gemelle potrebbero essere perturbati da un 'rumore' ε_{ij} che varia, da gemella a gemella (i), e di giorno in giorno (j):

$$\text{peso} = \mu + \beta_i + \varepsilon_{ij}$$

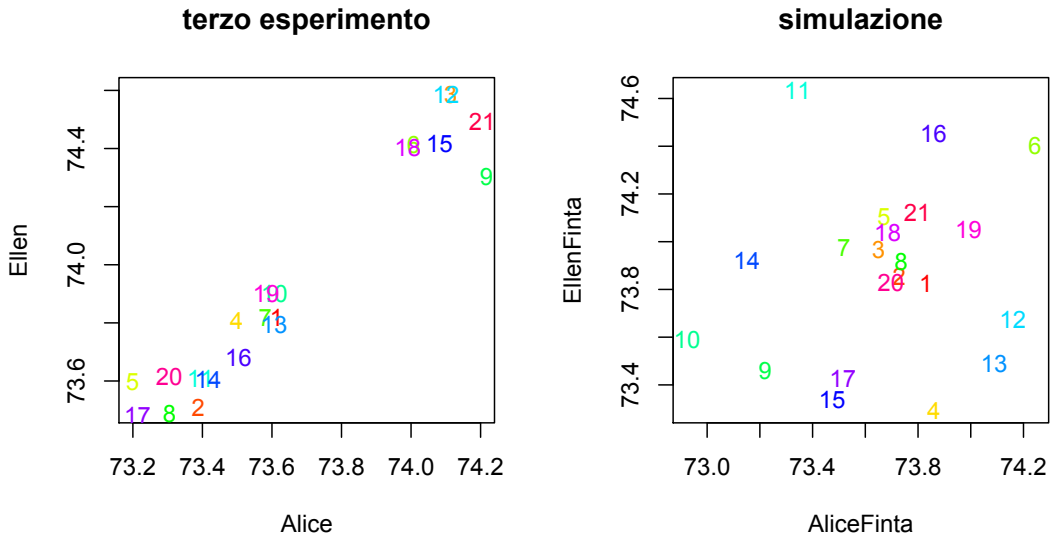
I software riescono a stimare, matematicamente, il comportamento casuale del 'rumore' ε_{ij}, indicando la quantità che si chiama *residual standard error*. Vediamolo con i comandi di R:

```
> gemelle21 = read.csv( file.choose(), header = TRUE)
> attach(gemelle21)
> modelloeffettifissi = lm( peso ~ gemella )
> summary(modelloeffettifissi)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.66190    0.08078 911.915   <2e-16 ***
gemellaellen  0.28095    0.11424   2.459   0.0183 *

Residual standard error: 0.3702 on 40 degrees of freedom
Multiple R-squared:  0.1314,Adjusted R-squared:  0.1096
F-statistic: 6.049 on 1 and 40 DF,  p-value: 0.01834
```

L'errore standard dei residui, `Residual standard error:  0.3702`, significa che il 'rumore' ε_{ij} è un numero casuale, distribuito in maniera gaussiana, con media zero e con deviazione standard 0.37 circa. Ma quello che non va bene è il fatto che l'errore standard dei residui venga determinato rispetto a 40 gradi di libertà, il che è un assurdo essendo che le nostre gemelle si sono pesate, ciascuna, 21 volte. Una simulazione mette in luce chiaramente il problema.

9.3 Una simulazione ci fa scoprire il colpevole

Riguardiamo con attenzione i dati grezzi del dataset `gemelle21` che abbiamo trascritto nella pagina precedente e notiamo che essi hanno un comportamento 'comune'. Per esempio, il terzo giorno Alice è andata a cena fuori (con Ellen) e ha mangiato un pochino di più del solito (anche Ellen), e la bilancia impietosa se ne è accorta. Ma una leggera dieta per Alice durante il quarto giorno (anche per Ellen) riporta le cose a posto al quarto giorno. Cosa vuol dire tutto questo? Vuol dire che ci appare del tutto logico che i dati di Ellen siano correlati con i dati di Alice; e sappiamo che in statistica la 'correlazione' si manifesta graficamente sul piano cartesiano con una nube di punti 'ordinata', come vediamo nel pannello sinistro della prossima figura:



Per intenderci, ricordiamo che al 21-esimo giorno i pesi di Alice ed Ellen erano rispettivamente 74.20 e 74.50; lo vediamo evidenziato nell'angolo in alto a destra del primo pannello. Il secondo pannello invece mostra una simulazione casuale ottenuta partendo dai parametri stimati dal `modelloeffettimisti` che abbiamo trovato. Come si vede, la nube di punti è del tutto caotica:

> *la simulazione ottenuta a destra mediante il modello statistico non rappresenta il fenomeno sperimentale di sinistra: pertanto il modello statistico ad effetti fissi non è* ***adeguato*** *alla realtà.*

Anche se proviamo a ripetere millanta volte questa simulazione (con il comando `simulate`) otterremo sempre una situazione disordinata di questo genere, e praticamente mai una come quella di sinistra. A sinistra, ci troviamo in una situazione di elevata **informazione**; a destra, in una situazione di assenza di informazione, ovvero di elevata **entropia** [3]. E questo contravviene alla richiesta di **adeguatezza** di un modello statistico [4, 9], che potremmo in maniera naïve esprimere in questo modo:

> *Un modello statistico M è adeguato a descrivere i dati D osservati a priori, rispetto ad un modello peggiore $\hat{M}$, se, generando a posteriori in maniera casuale per mezzo del modello M nuovi dati $D|M$, questi ultimi abbiano una 'grande' verosimiglianza; ovvero, la probabilità $P(D|M)$ che questi ultimi 'assomiglino' a quelli osservati sia 'molto elevata', rispetto a $P(D|\hat{M})$.*

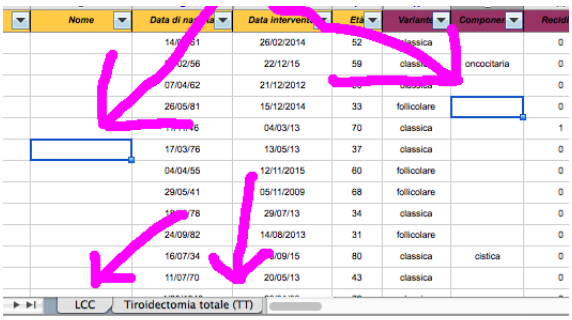
9.4 La proposta risolutiva

Attualmente, disponiamo di ottime soluzioni per fornire modelli statistici che gestiscano questo (ed altri!) tipi di difficoltà. Si chiamano **modelli ad effetti misti** e per avere un'idea sull'argomento vi consiglierei da dare un'occhiata al video che abbiamo caricato su YouTube sul nostro canale del Dipartimento di Scienze Mediche, Chirurgiche e della Salute:

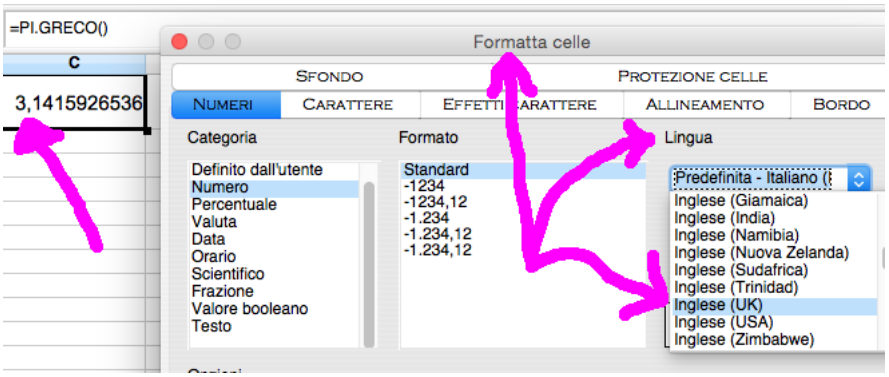
`https://www.youtube.com/watch?v=AH68lw10O4I`

10 Aiuto! Come si importa un dataset in R?

Dunque, innanzitutto il dataset deve essere molto ordinato, quanto spartano. Per esempio, quello propostoci inizialmente da Marina e Simone, è certamente un ottimo foglio di calcolo; ma non è un dataset:

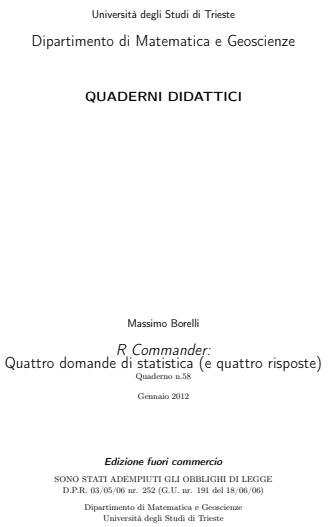


In un dataset non ci devono essere 'buchi bianchi', celle vuote. Se non disponiamo delle informazioni necessarie, dobbiamo usare una codifica per questo. R utilizza il codice `NA`, acronimo di Not Available. Nei fogli elettronici, in teoria, esiste la funzione NON.DISP() che genera il codice #N/D, ma viene usato da pochissime persone. Inoltre, se un'informazione è rilevante, essa deve apparire esplicitamente nel dataset. Per esempio, se vogliamo raffrontare il trattamento LCC versus il trattamento TT, essi non devono venir raccolti in due fogli separati, ma nel medesimo dataset, inserendo ad esempio una colonna denominata `trattamento`, che è appunto un fattore / variabile esplicativa. Abbiamo ancora una difficoltà: nel mondo anglosassone il punto decimale governa il mondo, mentre noi codifichiamo ancora questa informazione con la virgola. Mah! Per aggirare l'ostacolo, possiamo per esempio usare Open Office Calc, selezionare la cella in cui abbiamo il numero decimale scritto all'italiana ($\pi = 3,1415..$), selezionare Formato Celle e modificare la Lingua da quella predefinita (Italiano) in Inglese



11 Mah! Con questo R mi sembra tutto così difficile

Anche se io non sono un grande sostenitore di questo, vi devo confessare che esistono delle interfacce grafiche che vi aiutano a condurre l'analisi dei dati. Una delle iù celebri si chiama **R Commander**, ed è semplicissima da installare in un PC (molto meno in un Mac; non ne ho idea in Linux, perdonatemi!). Se proprio proprio vi sentite deboli e pasticcioni, e volete rimandare ad un - futuro, improbabile - domani il giorno in cui da pulcini della statistica vorrete diventare aquile, allora potete consultare la mia dispensuccola *R Commander:* Quattro domande di Statistica (e quattro risposte)


Università degli Studi di Trieste

Dipartimento di Matematica e Geoscienze

QUADERNI DIDATTICI

Massimo Borelli

R Commander:
Quattro domande di statistica (e quattro risposte)
Quaderno n.58

Gennaio 2012

Edizione fuori commercio

SONO STATI ADEMPIUTI GLI OBBLIGHI DI LEGGE
D.P.R. 03/05/06 nr. 252 (G.U. nr. 191 del 18/06/06)

Dipartimento di Matematica e Geoscienze
Università degli Studi di Trieste

che si scarica dal sito del Dipartimento di Matematica e Geoscienze, all'indirizzo:

`http://www.dmi.units.it/?q=node/635/d/2012`

Riferimenti bibliografici

[1] Naomi Altman and Martin Krzywinski. Points of significance: Sources of variation. *Nature methods*, 12(1):5–6, 2015.

[2] Göran Broström. *Event History Analysis with R.* CRC Press, 2012.

[3] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, 2003.

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

[5] Douglas Curran-Everett and Dale J Benos. Guidelines for reporting statistics in journals published by the american physiological society. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 287(2):G307–G309, 2004.

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

[7] Ruth Gilbert, Georgia Salanti, Melissa Harden, and Sarah See. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International journal of epidemiology*, 34(4):874–887, 2005.

[8] Srinivas Kondalsamy-Chennakesavan, Andreas Hackethal, David Bowtell, Andreas Obermair, Australian Ovarian Cancer Study Group, et al. Differentiating stage 1 epithelial ovarian cancer from benign ovarian tumours using a combination of tumour markers he4, ca125, and cea and patient's age. *Gynecologic oncology*, 129(3):467–471, 2013.

[9] John K. Kruschke. *Doing Bayesian data analysis: A tutorial with R, and BUGS.* Academic Press, 2011.

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

[11] Ronald L Wasserstein and Nicole A Lazar. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, (just-accepted):00–00, 2016.