Università degli Studi di Trieste

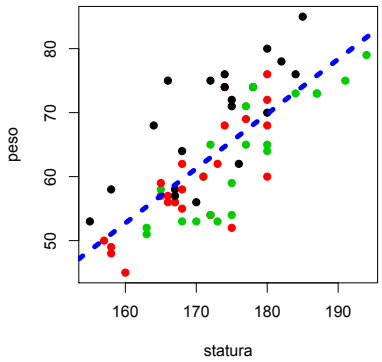
Corso di Laurea Magistrale
in Biotecnologie Mediche

Dispensa Didattica

603SM – Biostatistica

La retta di regressione, il modello lineare, il modello lineare generalizzato

www.dmi.units.it/borelli

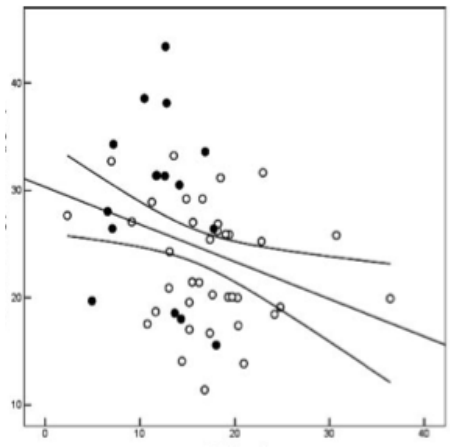


Autore
Massimo Borelli, Ph.D.

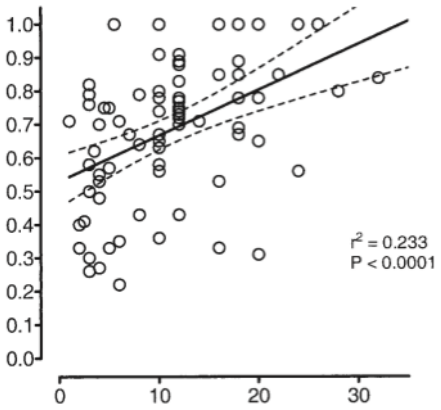
Anno Accademico 2016 – 2017

Indice

1 Motivazioni



Perseghin et al. [8] studiano la correlazione tra una determinata concentrazione serica e la sensibilità all'insulina in due gruppi di soggetti, rappresentati con punti di colorazioni opposte. Nel grafico si riporta una retta di regressione. Per quali motivo non sono state riportate due rette, l'una per i soggetti con storia di diabete e l'altra per i controlli? Forse quella retta di regressione è adeguata a descrivere entrambi i campioni?



Nel paper Naguib et al.[7] una nube di punti viene modellata per mezzo di una retta di regressione che appare essere caratterizzata da un p-value altamente significativo. La nube però appare caotica ed il coefficiente di determinazione $R^2 = 0.233$ è piuttosto basso. Vi sono però quantità di punti che appaiono tra loro 'allineati'. Quali sono gli strumenti che consentono di decidere se il modello regressivo utilizzato dagli autori è adeguato a descrivere i loro dati sperimentali?

2 La retta di regressione

2.1 Il dataset

Facciamo riferimento ad un insieme di dati che abbiamo raccolto intervistando una coorte di studenti del primo anno di medicina:

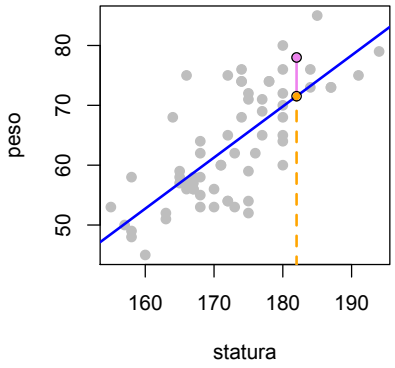
```
www <- "http://www.dmi.units.it/~borelli/dataset/studentiannoscorso.txt"
studentimedicina <- read.table( www , header = TRUE )
attach(studentimedicina)
```

	anno	genere	statura	peso	scarpe	fumo	sport
1	1987	f	155	53	36	NO	poco
2	1987	f	157	50	37	NO	saltuario
3	1989	f	158	48	36	NO	saltuario
4	1987	f	158	49	36	NO	saltuario
..	..	..	..	..	..	..	..
64	1989	m	191	75	44	NO	tanto
65	1989	m	194	79	46	SI	tanto

2.2 L'idea generale

Ci proponiamo di verificare se la `statura` possa essere, in senso statistico, un predittore del `peso`. In termini 'scolastici', trovare una retta di regressione significa cercare un opportuno coefficiente angolare m ed un'opportuna intercetta (o quota) q in modo tale che la retta $y = m \cdot x + q$ attraversi una nuvola di punti disegnata sul piano cartesiano xy nel 'miglior modo possibile'.

Tale 'miglior modo possibile' -usualmente- si ottiene per mezzo di un procedimento detto *ai minimi quadrati*; si cerca cioè di minimizzare, complessivamente, le distanze dei **residui** (come per esempio il segmentino viola nel grafico della pagina successiva), ossia gli 'scostamenti' da ciascun punto della nube di coordinate (x_i, y_i) rispetto alla sua proiezione verticale (in colore arancione) sulla retta di regressione $y = m \cdot x + q$. In altri termini, per ogni x_i si definisce come residuo ϵ_i la distanza che intercorre tra la ascissa del punto, y_i, e la sua proiezione sulla retta, $m \cdot x_i + q$.



Con R mediante il comando `lm`, oltre a trovare immediatamente tali m e q, abbiamo anche una quantità di strumenti che ci consente di valutare l'affidabilità e la 'qualità' del modello che noi abbiamo formulato. Iniziamo a vedere come vanno le cose se vogliamo ipotizzare che la `statura` del dataset studentiannoscorso.txt sia un predittore del loro `peso`.

```
modello <- lm (peso ~ statura)
plot (statura, peso)
abline(modello, lty = 3)
```

Con il termine `modello` abbiamo 'battezzato' la retta di regressione, e l'abbiamo tracciata con il comando `abline`, in maniera tratteggiata (`lty = 3`). Vediamo i coefficienti della retta di regressione, digitando semplicemente `modello`. Ecco qui di seguito riportato l'output del software:

```
Call:
lm(formula = peso ~ statura)

Coefficients:
(Intercept)      statura
    -83.891        0.854
```

Cosa significa? Abbiamo determinato una stima puntuale della intercetta q e del coefficiente angolare m. In parole povere, il `peso` del nostro studente i-esimo è legato alla sua `statura` dalla relazione:

$$peso_i = 0.854 \cdot statura_i - 83.891 + \epsilon_i$$

ed ϵ_i è il **residuo** che viene attribuito a ciascun soggetto in maniera da poter giustificare quell'errore in più o in meno che il modello lineare non riesce a 'spiegare perfettamente'. La teoria matematica ci richiede delle particolari ipotesi su tali residui, che costituiscono la **componente aleatoria** del modello lineare: essi devono essere dei numeri casuali distribuiti normalmente, indipendenti tra loro, con media nulla (il che significa che non ci sono errori sistematici nelle misure che stiamo analizzando), e con una dispersione (deviazione standard) σ costante.

Dicevamo che R fornisce la possibilità di giudicare criticamente la adeguatezza della regressione, e questo si fa innanzitutto con il commando `summary`:

```
summary(modello)
```

Riportiamo qui di seguito l'output, che contiene moltissime informazioni, e che sarebbe opportuno riuscire ad essere in grado di interpretarle tutte, o quasi tutte, nella maniera corretta.

```
Call:
lm(formula = peso ~ statura)

Residuals:
    Min      1Q  Median      3Q     Max
-13.546  -4.209  -1.569   4.431  17.139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -83.89056   16.67708   -5.03 4.34e-06 ***
statura       0.85392    0.09649    8.85 1.18e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.459 on 63 degrees of freedom
Multiple R-squared: 0.5542,     Adjusted R-squared: 0.5471
F-statistic: 78.31 on 1 and 63 DF,  p-value: 1.18e-12
```

Per iniziare, osserviamo che la nostra nuvola di punti ha un **coefficiente di determinazione** R^2 (`Multiple R-squared`) pari a 0.5542, che non è vicino ad 1 come tutti

vorremmo, ma che non è neanche vicino allo 0, come nessuno vuole; ricordiamo che R^2 è il coefficiente di correlazione lineare di Pearson ρ elevato al quadrato:

```
summary(modello)$r.squared
cor(statura, peso)^2
```

Ci viene anche fornita una stima puntuale del valore di σ, la deviazione standard dei residui (`Residual standard error`, 6.459) modellati sulla variabile aleatoria $N(0, \sigma)$:

```
summary(modello)$sigma
```

Ciò che però colpisce sempre l'occhio del ricercatore sono le stelline, `***`. Nella nostra analisi, le stelline ci suggeriscono che la `statura` è un predittore altamente significativo del `peso`; nel senso geometrico che la retta di regressione ha una pendenza m e un'intercetta q 'vere e proprie', ossia entrambi i numeri sono diversi da zero in senso statistico: non può quindi darsi (con un grado di fiducia del 95 per cento, si intende) che uno di questi due parametri sia trascurabile, e che la retta di regressione possa ricondursi alla forma $y = mx$ (una retta che passa per l'origine degli assi), oppure $y = q$ (una retta orizzontale).

2.3 La diagnostica del modello lineare

In ogni modello di regressione dobbiamo prestare attenzione ai **punti separati** dalla nuvola, perché questi potrebbero avere una **forza di leva** tale da riuscire a influenzare i coefficienti della retta, riuscendo a 'spostarla'. E' un argomento che è stato messo benissimo in luce da F. Anscombe, con il suo *quartetto*:

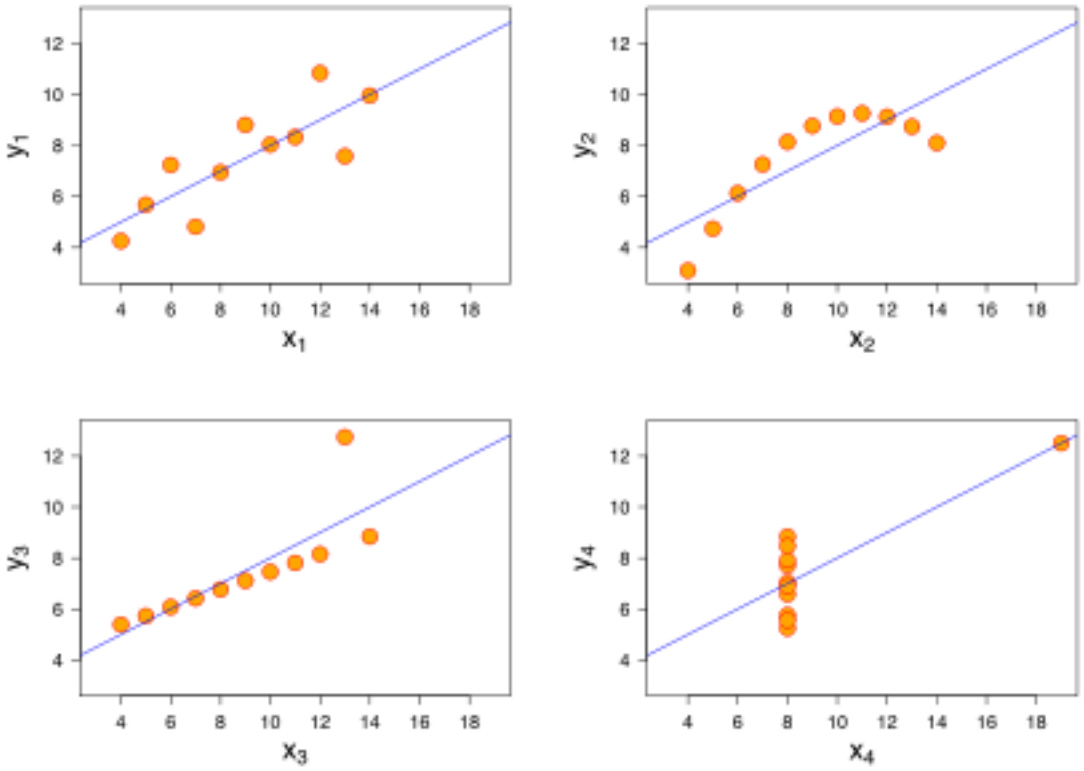


Figura 1: http://en.wikipedia.org/wiki/Anscombe's_quartet

Per esempio, consideriamo la studentessa numero 15:

```
genere[15]
statura[15]
peso[15]
```

Il nostro modello di regressione non distingue però tra maschi e femmine, e prevede genericamente che per un'altezza di 166 centimetri sia verosimile pesare attorno ai 58 chili:

```
0.8539245 * statura[15] - 83.89056
modello$fitted[[15]]
```

Pertanto, nel caso particolare della nostra studentessa, il modello ha un residuo ϵ_{15} che ci potrebbe apparire piuttosto elevato:

```
peso[15] - (0.8539245 * statura[15] - 83.89056)
modello$residual[[15]]
```

Ci chiediamo: se questa studentessa non fosse stata presente nel dataset, come sarebbe cambiata la retta di regressione? Di tanto o di poco?

```
modello_senza15 <- lm (peso ~ statura,
studentimedicina[c(1:14, 16:65),] )
modello_senza15
```

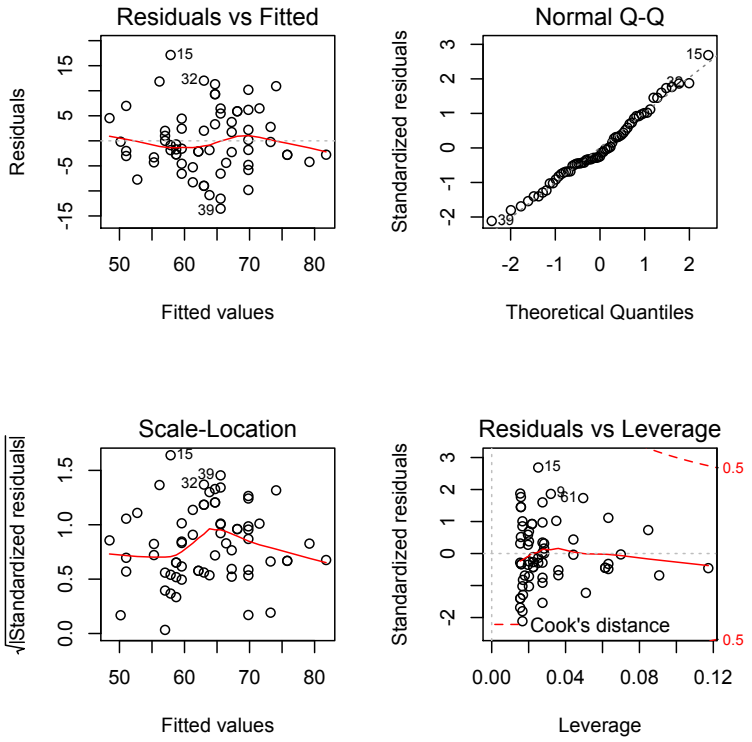
I coefficienti della regressione si sono modificati, ma non è facile giudicare in senso critico se lo spostamento sia di lieve o rilevante entità. E' proprio per questo motivo che sono state implementate in R numerose istruzioni per effettuare la diagnostica del modello lineare, che in pratica si riconduce ad esaminare visualmente questi quattro grafici:

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

Il primo grafico raffigura i residui ϵ_i. La retta grigia tratteggiata indica il valore teorico medio di zero (infatti, in teoria, la media degli errori dovrebbe essere nulla). La curva rossa è uno **smoother**, ossia una curva 'pesata localmente' che riassume l'andamento della nube dei residui: essa appare ragionevolmente piatta, e dunque non abbiamo a preoccuparci della ipotesi che la media sia pari a zero. La nube dei punti inoltre appare sufficientemente caotica (non ha cioè una forma di torta, o di cuneo, o qualche particolare curvatura), e questo ci fornisce un'indicazione affidabile del fatto che i residui abbiano deviazione standard σ costante. Il grafico però ci suggerisce di stare attenti agli studenti numero 1, numero 64 e numero 65 perché i loro residui potrebbero essere 'particolarmente elevati' rispetto agli altri.

Nel secondo pannello, il QQplot ci rassicura: non dobbiamo temere per la normalità dei residui. Il terzo pannello riflette quanto già visto nel primo -e lo trascuriamo-, mentre dal quarto grafico deduciamo che non ci sono punti isolati nella nube che abbiano una rilevante forza di leva sulla retta di regressione (e la distanza di Cook[1] è l'oggetto matematico adatto per trattare questo aspetto). Quindi la studentessa numero 15 era stata da noi 'ingiustamente accusata' di essere differente dal punto di vista morfometrico dai suoi colleghi.

[1] http://en.wikipedia.org/wiki/Cook's_distance



2.4 .. ma la parola *lineare* non vuol dire *rettilineo*

Facciamo un precisazione molto, molto importante: il termine modello **lineare** non sta a significare che la nuvola di punti debba necessariamente essere (approssimativamente) *rettilinea*. Il concetto di 'linearità' è inteso nella sua accezione matematica (senza essere troppo pesanti, una funzione è lineare ad esempio se $f(x + 2y) = f(x) + 2f(y)$), e riguarda i coefficienti della regressione: ad esempio, la parabola $y = ax^2 + bx + c$ non è una funzione lineare rispetto alla x, ma lo è rispetto ai coefficienti a, b e c.

Prendiamo ad esempio il dataset `decay` descritto da Crawley[**?**] a pagina 146:

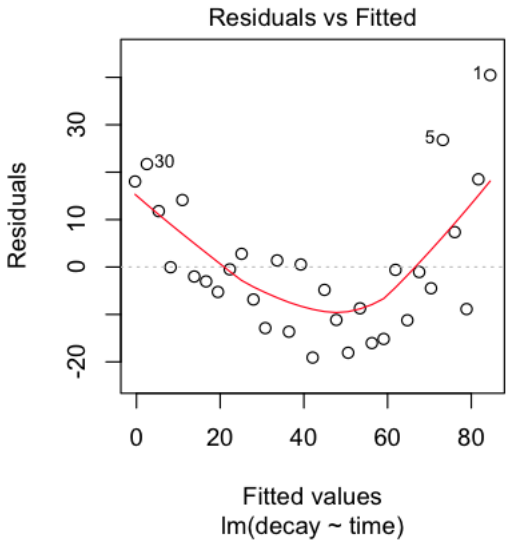
```
time = 0:30
decay = c(125.0, 100.2, 70.0, 83.4, 100.0, 65.9, 66.5, 53.5, 61.3,
43.9, 40.2, 44.7, 32.5, 36.6, 40.1, 23.0, 39.8, 22.8, 35.0, 17.9,
21.1, 27.9, 21.8, 14.2, 13.6, 11.8, 25.1, 8.1, 17.1, 24.2, 17.7)
```

Cercando una curva di calibrazione che descriva il fenomeno, si potrebbe provare con:

```
decadimento1 = lm(decay ~ time)
```

il cui `summary` fornirebbe una retta di regressione con entrambi coefficienti significativi:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	84.5278	5.0298	16.81	0.0000
time	-2.8290	0.2880	-9.82	0.0000



ma effettuando una diagnostica del modello ci si accorge che c'è una considerevole curvatura nei residui. Conviene dunque introdurre un termine quadratico nel modello:

```
decadimento2 = lm(decay ~ time + I(time^2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	106.3731	4.6578	22.84	0.0000
time	-7.3487	0.7187	-10.23	0.0000
I(time^2)	0.1507	0.0231	6.51	0.0000

ed effettuando nuovamente la diagnostica del modello si vede che il problema è stato parzialmente eliminato - come noto, il decadimento segue una legge esponenziale, e dunque una soluzione ottimale sarebbe quella di considerare una trasformazione logaritmica:

```
decadimento3 = lm( log(decay) ~ time + I(time^2))
```

2.5 .. ed attenzione ai p-value

Ritorniamo all'esempio del dataset `studentimedicina`. Abbiamo visto che con il

```
summary(modello)
```

ottenevamo dei p-value per l'intercetta q e per la pendenza m:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-83.8906	16.6771	-5.03	4.34e-06 ***
statura	0.8539	0.0965	8.85	1.18e-12 ***

Si tratta sempre di stime marginali che non tengono conto della correlazione esistente tra le due quantità, i.e. della correzione per la molteplicità. In Bretz et al. 2011 [**?**] si suggerisce di sfruttare il pacchetto `multcomp` per ottenere stime più attendibili, implementando dei metodi 'single-step' che tengono conto delle correlazioni tra i coefficienti della regressione:

```
library("multcomp")
aggiustato = glht(modello, linfct = diag(2))
summary(aggiustato)
```

In questo caso, come si vede, non c'è differenza apprezzabile:

	Estimate	Std. Error	t value	Pr(>|t|)
1 == 0	-83.89056	16.67708	-5.03	4.71e-06 ***
2 == 0	0.8539	0.0965	8.85	< 1e-10 ***

3 La Ancova

3.1 A cosa serve

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 [5] 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 mi sento di condividere – in linea di principio – 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 2: covariate o predittori?

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

3.2 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`:

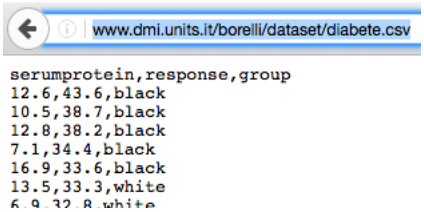
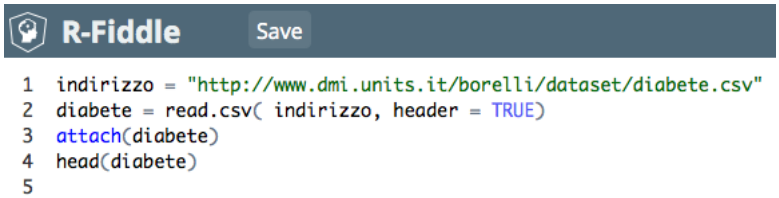


Figura 3: il dataset `diabete.csv` on line sul mio sito web

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:

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 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.

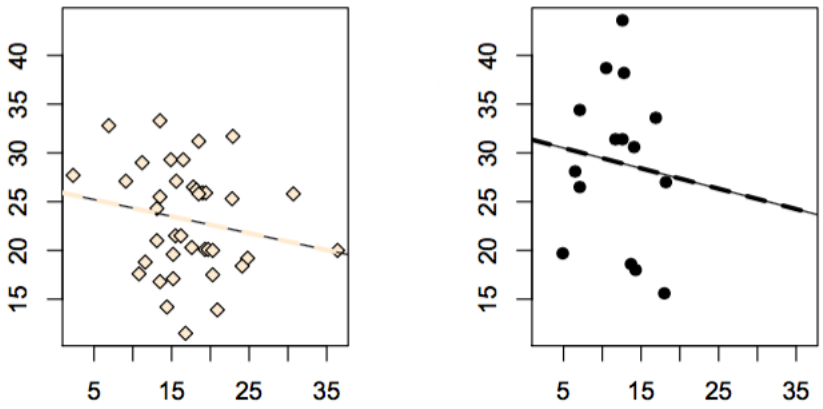
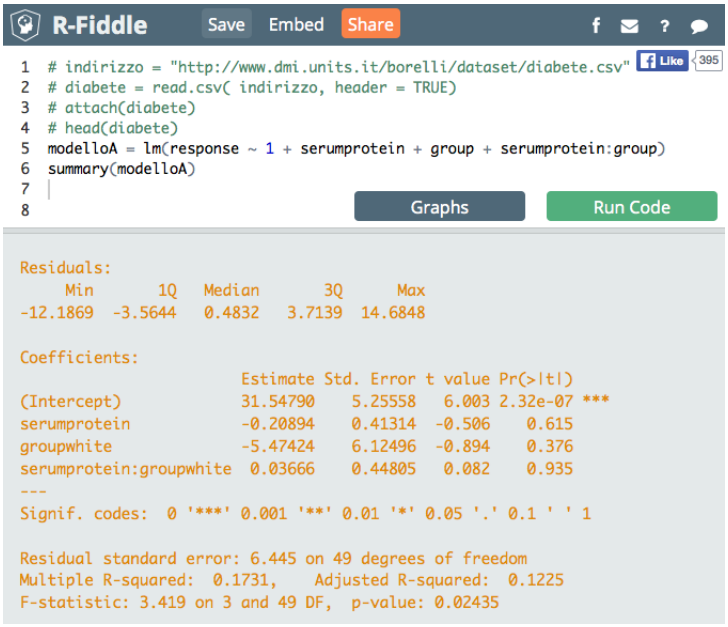


Figura 4: il modello massimale: due rette di regressione, entrambe con differenti pendenze e differenti intercette.

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:

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

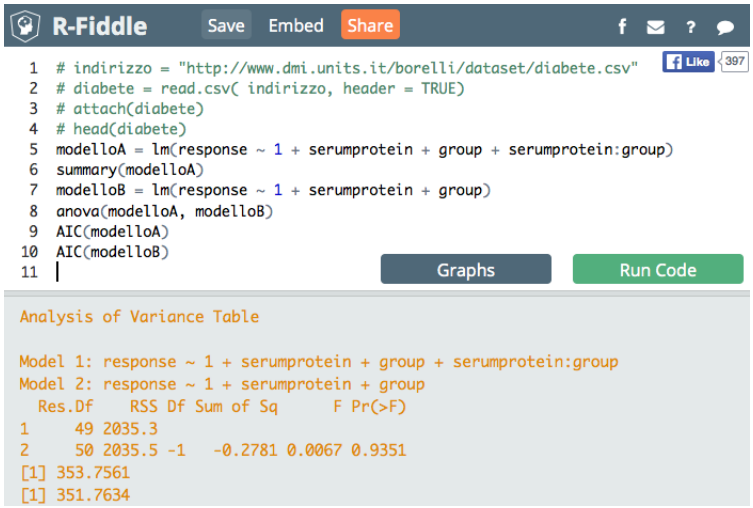
mentre la retta dei neri ha equazione:

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

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`.

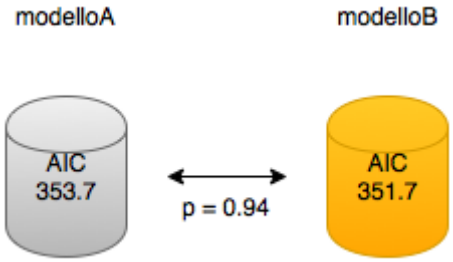
Per fare questo controllo, impostiamo un `modelloB` additivo e facciamo due verifiche: un'**analisi della devianza** con il comando `anova` e un'analisi dei **criteri di informazione** con il comando `AIC`:

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



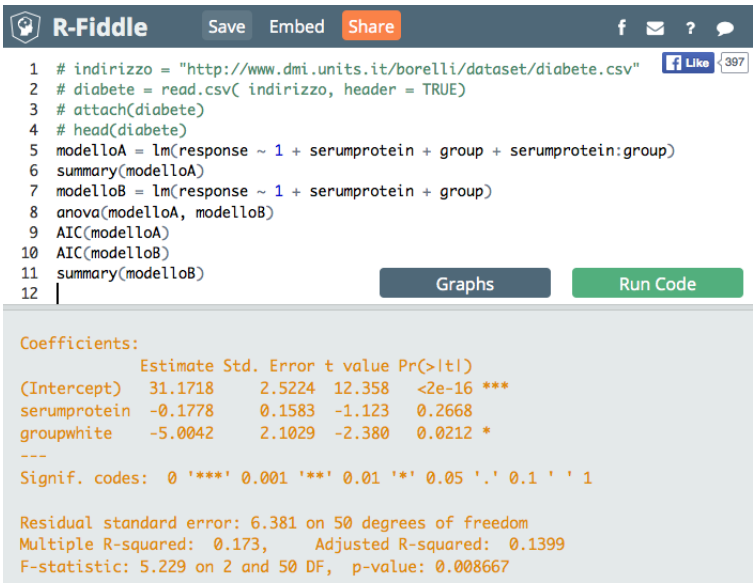
Leggete su Wikipedia chi è stato il professot Hirotsugu Akaike, e cosa significa il suo criterio di informazione, che funziona all'insegna di *'small is beautiful'*, 'nelle botti piccole ci sta il vino buono'. Siccome il `modelloB` ha un criterio di informazione inferiore a quello del `modelloA`, e siccome 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)



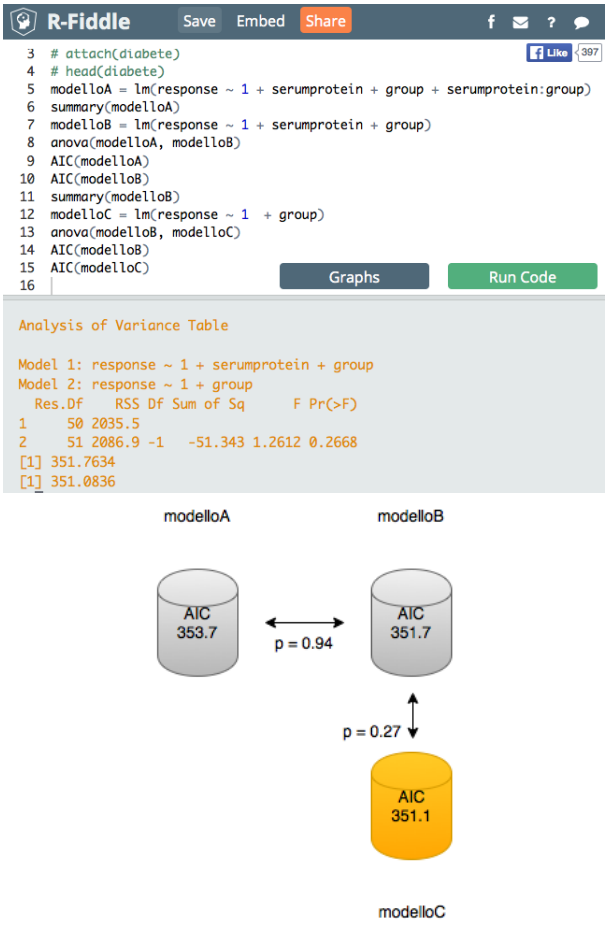
Esaminiamo tuttavia il `summary` del `modelloB`:

```
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 ricordate 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)

3.3 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

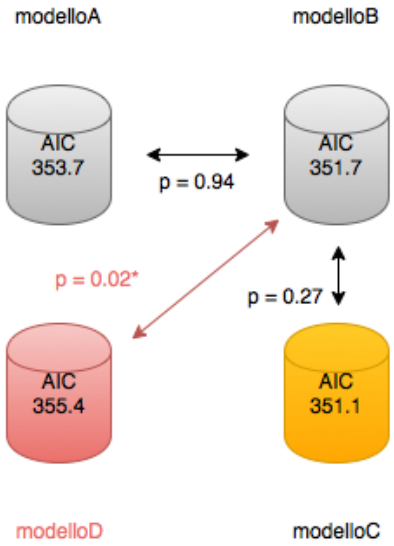
3.4 Quali sono gli errori da evitare

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, $\sigma_\varepsilon = 6.66$.

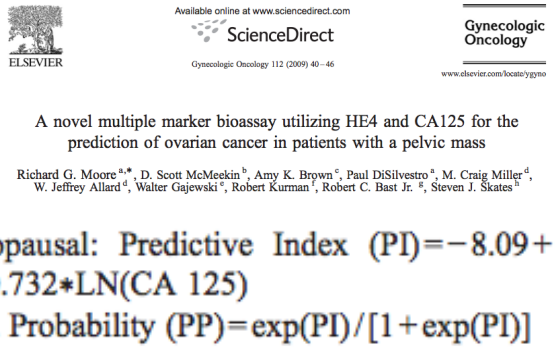


E perciò a nessuno, e dico nessuno, dovrebbe mai venire in mente di piazzare il `modelloD` in una tesi di laurea. Men che meno, in un paper. E ho detto tutto!

4 La regressione logistica

4.1 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**

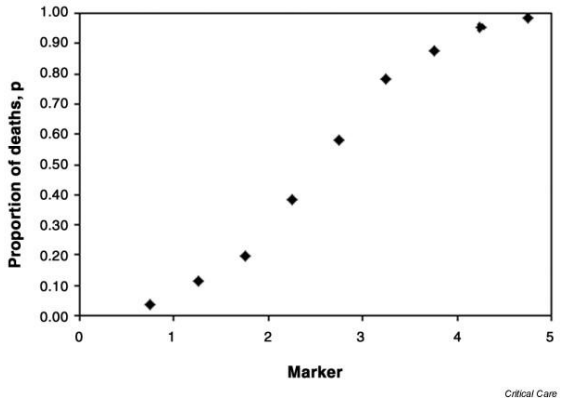


Figura 5: Logistic regression, http://ccforum.com/content/9/1/112

- 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'`)

4.2 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
```

4.3 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).

4.4 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 [4] 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'.

5 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:

5.1 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

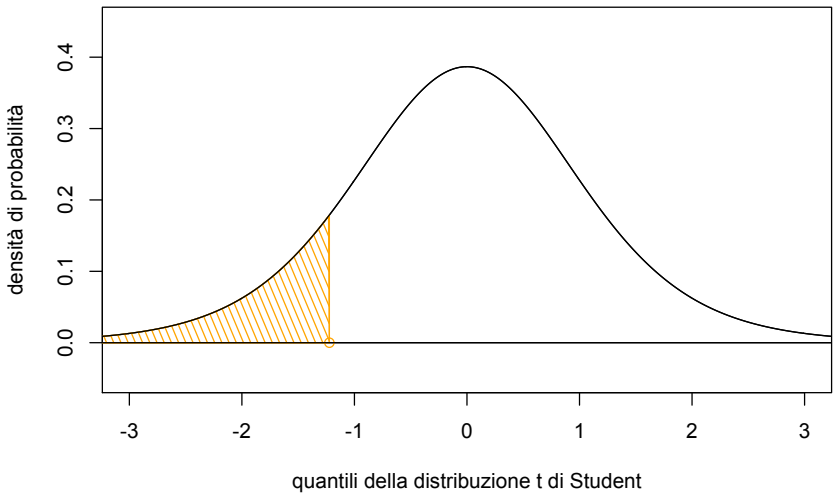
Per dirimere la questione esse ricorrono al celebre test t di Student. Come tutti ricordano, si vuole decidere se la media dei pesi di Alice sia diversa 'in senso statistico' dalla media dei pesi di Ellen, immaginando che per ciascuna di esse siano stati osservati cinque numeri casuali provenienti da 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?

5.2 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** [9], 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):

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

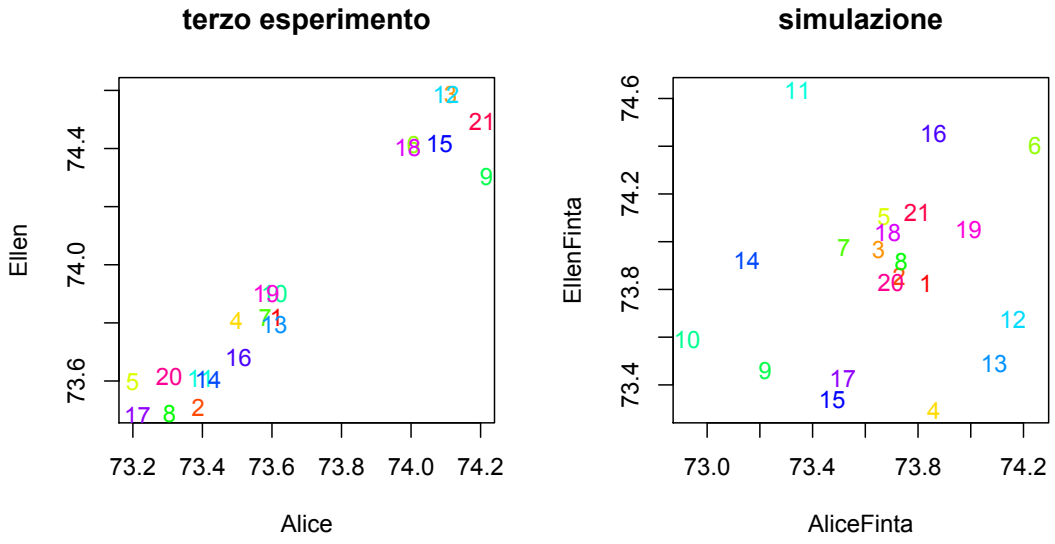
I software riescono a stimare, matematicamente, il comportamento casuale del 'rumore' ε_{ij}, indicando la quantità che si chiama *residual standard error*. Vediamolo con i comandi di R:

```
> 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.

5.3 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** [2]. E questo contravviene alla richiesta di **adeguatezza** di un modello statistico [3, 6], 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 $P(D|\hat{M})$.*

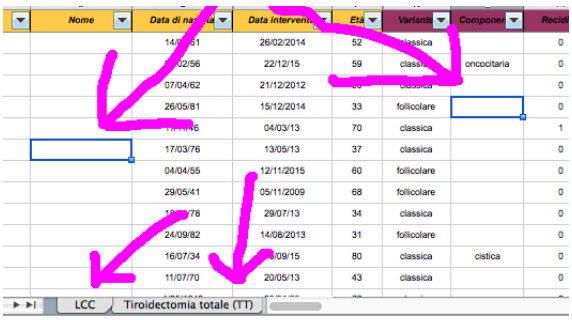
5.4 La proposta risolutiva

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:

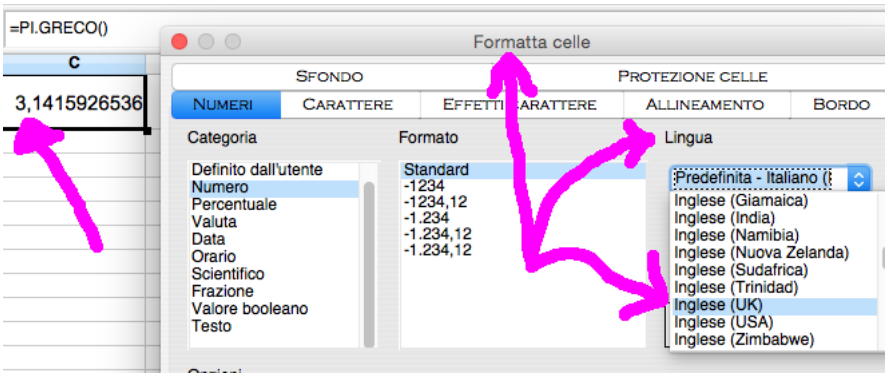
`https://www.youtube.com/watch?v=AH68lw10O4I`

6 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:

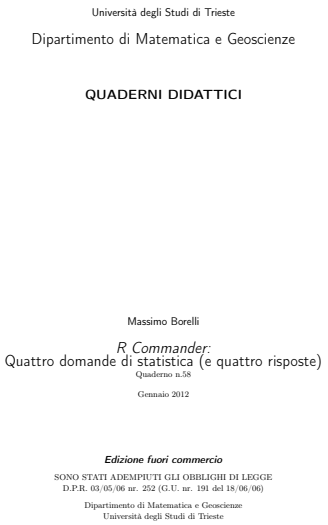


In un dataset non ci devono essere 'buchi bianchi', celle vuote. Se non disponiamo delle informazioni necessarie, dobbiamo usare una codifica per questo. R utilizza il codice `NA`, acronimo di Not Available. Nei fogli elettronici, in teoria, esiste la funzione NON.DISP() che genera il codice #N/D, ma 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



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

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

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

[5] 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.

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

[7] Mohamed Naguib, Cynthia A Lien, John Aker, and Rudolfo Eliazo. Posttetanic potentiation and fade in the response to tetanic and train-of-four stimulation during succinylcholine-induced block. *Anesthesia & Analgesia*, 98(6):1686–1691, 2004.

[8] Gianluca Perseghin, Guido Lattuada, Francesco De Cobelli, Antonio Esposito, Elena Belloni, Tamara Canu, Francesca Ragogna, Paola Scifo, Alessandro Del Maschio, and Livio Luzi. Serum retinol-binding protein-4, leptin, and adiponectin concentrations are related to ectopic fat accumulation. *The Journal of Clinical Endocrinology & Metabolism*, 92(12):4883–4888, 2007.

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