Sixth Torino R net meeting - Campus Luigi Einaudi, Torino, November 21,2013

Predittori e risposte: il pane quotidiano del consulente statistico

Mauro Gasparini Professor of Statistics, Editor of *Biometrical Journal*
Dipartimento di Scienze Matematiche
Politecnico di Torino
gasparini@calvino.polito.it
http://calvino.polito.it/~gasparini

Predittori e risposte nella consulenza statistica

In un progetto statistico, tipicamente leghiamo
il comportamento probabilistico
di una (o più) variabili risposta (o semplicemente risposta)
a una o più predittori per:

descrivere

spiegare

predire

decidere

L'istruzione più utile in R

```
risposta ~ predittori
```

Esempio più noto: i modelli lineari `lm()`.

Una risposta quantitativa, un predittore binario

Esempio con `lm`:

```
scarpe <- read.table("scarpe.txt",header=T)
summary(lm(consumo~tipo, data=scarpe))
```

In alternativa:

```
t.test(scarpe[scarpe$tipo=="A","consumo"],
       scarpe[scarpe$tipo=="B","consumo"],
       var.equal=T)
```

	consumo	tipo
1	13.2	A
2	8.2	A
3	10.9	A
4	14.3	A
5	10.7	A
6	6.6	A
7	9.5	A
8	10.8	A
9	8.8	A
10	13.3	A
11	14.0	B
12	8.8	B
13	11.2	B
14	14.2	B
15	11.8	B
16	6.4	B
17	9.8	B
18	11.3	B
19	9.3	B
20	13.6	B

Una risposta quantitativa, due predittori categorici

Esempio con `lm`:

```
attach(colla <- read.table("colla.txt",header=T))
colla$tipocolla <- as.factor(colla$tipocolla)
# lm e confronto tra modelli (test F)
additivo <- lm(pressione ~ legno + tipocolla)
grande <- lm(pressione ~ legno * tipocolla)
anova(additivo,grande); detach(colla)
```

In alternativa:

```
# in alternativa, confronto tra modelli stile aov
additivo_aov <- aov(pressione ~ legno + tipocolla)
grande_aov <- aov(pressione ~ legno * tipocolla)
anova(additivo_aov, grande_aov)
# sul concetto di interazione
interaction.plot(legno, tipocolla, pressione)
# confronti multipli tra medie con Tukey
par(mfrow=c(3,1));  plot(TukeyHSD(grande))
```

	legno	tipocolla	pressione
1	a	1	196
2	a	1	208
3	a	1	247
4	a	1	216
5	a	1	221
	.....		
	.....		
14	a	3	248
15	a	3	272
16	b	1	216
17	b	1	228
18	b	1	240
	.....		
	.....		
40	c	2	222
41	c	3	255
42	c	3	251
43	c	3	261
44	c	3	258
45	c	3	247

Una risposta quantitativa, predittori misti

Esempio di consumo di calore per abitazione in funzione della temperatura esterna e di un intervento di coibentazione:

```
coibent<-read.table("coibentazione.txt",header=T)
attach(coibent); plot(temp,consumo,type="n")
points(temp[quando=="prima"],
       consumo[quando=="prima"],pch=1)
points(temp[quando=="dopo"],
       consumo[quando=="dopo"],pch=2,col=2)
```

Modello lineare con predittori misti (ANCOVA):

```
### relazione inversa consumo-temperatura
summary( lm(consumo ~ temp)
### due rette parallele
summary( lm(consumo ~ quando + temp)
detach(coibent)
```

	quando	temp	consumo
1	prima	-0.8	7.2
2	prima	-0.7	6.9
3	prima	0.4	6.4
4	prima	2.5	6.0
5	prima	2.9	5.8
6	prima	3.2	5.8
7	prima	3.6	5.6
8	prima	3.9	4.7
9	prima	4.2	5.8
10	prima	4.3	5.2
11	prima	5.4	4.9
12	prima	6.0	4.9
...	...	...	...
...	...	...	...
48	dopo	5.3	3.7
49	dopo	6.2	2.8
50	dopo	7.1	3.0
51	dopo	7.2	2.8
52	dopo	7.5	2.6
53	dopo	8.0	2.7
54	dopo	8.7	2.8
55	dopo	8.8	1.3
56	dopo	9.7	1.5

Inquadramento storico

La unificazione di t-test, ANOVA, ANCOVA, regressione semplice e regressione multipla in un unico inquadramento logico è stato uno dei grandi risultati del dopoguerra.

Più recentemente (dagli anni '70 in poi), le stesse tecniche sono state estese a riposte diverse da una variabile quantitativa come:

- conteggi,
- risposte binarie,
- risposte categoriche,
- risposte di durata,
- risposte multivariate
- serie temporali
- dati longitudinali.

A questo dobbiamo aggiungere il contributo fondamentale dato dall'introduzione degli effetti casuali nell'arsenale statistico...

... e la ovvia connessione con la statistica bayesiana.

La mia consulenza statistica

Prove cliniche per pharma (big e small)

Prove cliniche indipendenti

Ricerche accademiche in biostatistica

Esperimenti industriali per manifattura

Generalizzazione di `lm()` in `glm()`

Un primo grosso contributo è stato naturalmente l'invenzione dei modelli lineari generalizzati e la loro codifica in R

```
glm(formula,                    ### la nostra "risposta ~ predittori"
data,                           ### database interessato
subset,                         ### seleziona solo una parte del db
na.action,                      ### cosa fare con i dati mancanti
offset,                         ### predittore fisso
family = gaussian,              ### famiglia esponenziale, gaussiana default
start = NULL,                   ### valori iniziali per max verosim numerica
control = glm.control(...),     ### come gestire la numerica per i GLM
model = TRUE,                   ### configurazione output: vuoi model.frame?
y = TRUE,                       ### configurazione output: vuoi la y?
x = FALSE, ...)                 ### configurazione output: vuoi la x?
```

I benefici della programmazione a oggetti

I principali metodi e funzioni accessorie di `glm()` sono ereditate da `lm()`:

- `summary()` sintesi dei risultati
- `coef()` estrazione dei coefficienti stimati
- `confint()` calcolo degli intervalli di confidenza
- `residuals()` calcolo dei residui
- `anova()` confronto tra modelli annidati
- `plot()` grafici diagnostici
- `predict()` predizione e previsione

Una risposta binaria: Esempio TAGLIO

Si tratta di valutare/predire il rischio di taglio cesareo in emergenza (durante il parto, non programmato).

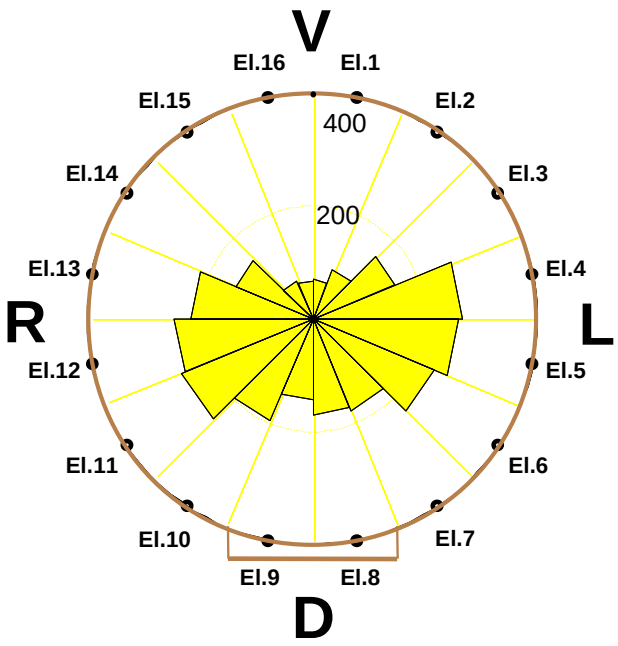
```
studio <- read.table("taglio-dati.txt", header=T, na.strings=".")
attach(studio)
head(studio) # un'occhiata ai dati
str(studio)

# risp=0,1 (taglio cesareo sì o no)
# età materna(eta),
# altezza materna (alt),
# peso materno (peso),
# Body Mass Index(bmi),
# primipara? (parita),
# età gestionale(eg),
# aumento di peso in gravidanza(ip),
# livello della parte presentata all'inizio travaglio(pp)
# allarme da Bishop score? (bs)
```

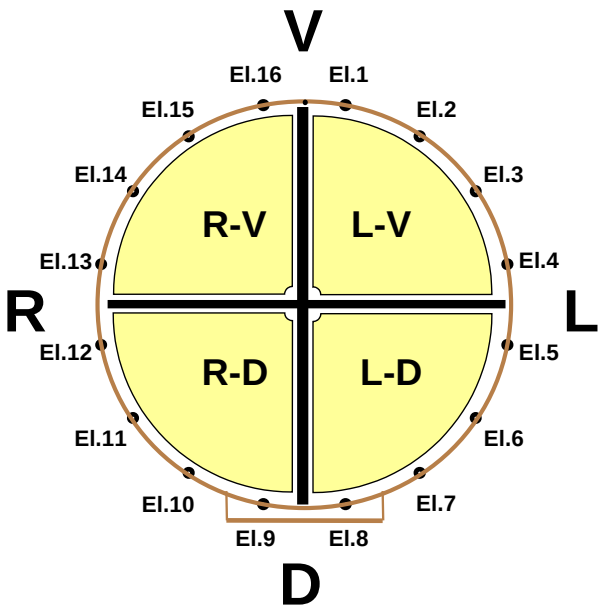
Una risposta binaria, vari predittori: regressione logistica

```
# fittiamo un GLM per dati binari con il link logistico (binomial)
mod1 <- glm(risp ~ alt+peso+eta+parita+eg+bs, family = binomial)
summary(mod1)

# confronto tra modelli
# anche con glm() c'é anova ma meno formale:
mod2<- glm(risp ~ alt + eta + parita + eg, family = binomial)

# a differenza di lm(), glm() è riluttante a dare un p-value
# si vede comunque che una devianza di 1.8059 su 2 df non è significativa
# (giustificando la restizione al modello piccolo mod2)
anova(mod2,mod1)
# come modello finale ci si puo' limitare a mod2
```

Una risposta di conteggio: Esempio IZ

Un altro esempio ostetrico (dettagli omessi): distribuzione circolare empirica della variabile `iz`, zone di innervazione di un muscolo circolare.
Due posizioni (`ventral/dorsal`)
e un lato (`left/right`) sono due fattori o meglio, la loro combinazione `quad` uno solo?





The statistical database format

	pat	weight	age	iz	arv	bbweight	type	time	pos	side	quad
1	1	78	31	0	2.66	NA	1	bef	ventr	left	vl
2	1	78	31	2	2.66	NA	1	bef	dors	left	dl
3	1	78	31	0	2.59	NA	1	bef	ventr	right	vr
4	1	78	31	0	2.72	NA	1	bef	dors	right	dr
5	1	78	31	0	4.21	NA	1	aft	ventr	left	vl
6	1	78	31	3	5.92	NA	1	aft	dors	left	dl
7	1	78	31	0	5.89	NA	1	aft	ventr	right	vr
8	1	78	31	4	6.58	NA	1	aft	dors	right	dr
.....											
23	3	60	23	1	7.64	3836	4	aft	ventr	right	vr
24	3	60	23	3	10.24	3836	4	aft	dors	right	dr
25	4	70	34	0	5.48	3230	1	bef	ventr	left	vl
26	4	70	34	5	8.77	3230	1	bef	dors	left	dl

Risposte, predittori ed effetti casuali

La risposta `iz` è un conteggio, quindi un intero. Possibili predittori di interesse sono:

- caratteristiche demografiche del paziente
- diversi fattori categorici:

  `type (1=Caesarean section 4=episiotomy),`

  `time (before/after delivery),`

  `quad (vl/vr/dl/dr)`
- le loro interazioni.

Il fatto nuovo è la comparsa dell'identità dal paziente che necessita di un effetto casuale, a differenza degli effetti fissi di interesse come il tempo `time=after` versus `time=before` e suoi contrasti, oppure il quadrante eccetera.

Software for random effects

The `lme4` package in `R`.

The `PROC MIXED` procedure in `SAS`.

Other specialized software, like `NONMEM` and the more recent `Monolix` for PK/PD.

For non-gaussian responses, linear mixed models have been generalized to generalized linear mixed models.

Analysis of our data using `lme4`

Call the `lme4` package in R ....

```
library(lme4)
```

... then fit a generalized linear mixed effect model with three fully crossed fixed predictors only one random effect for patient, using data `dati14` and the poisson model for the count response variable `iz`:

```
grande <- glmer(iz~type*time*quad+(1|pat), data=dati14, family=poisson)
```

Significance of interactions

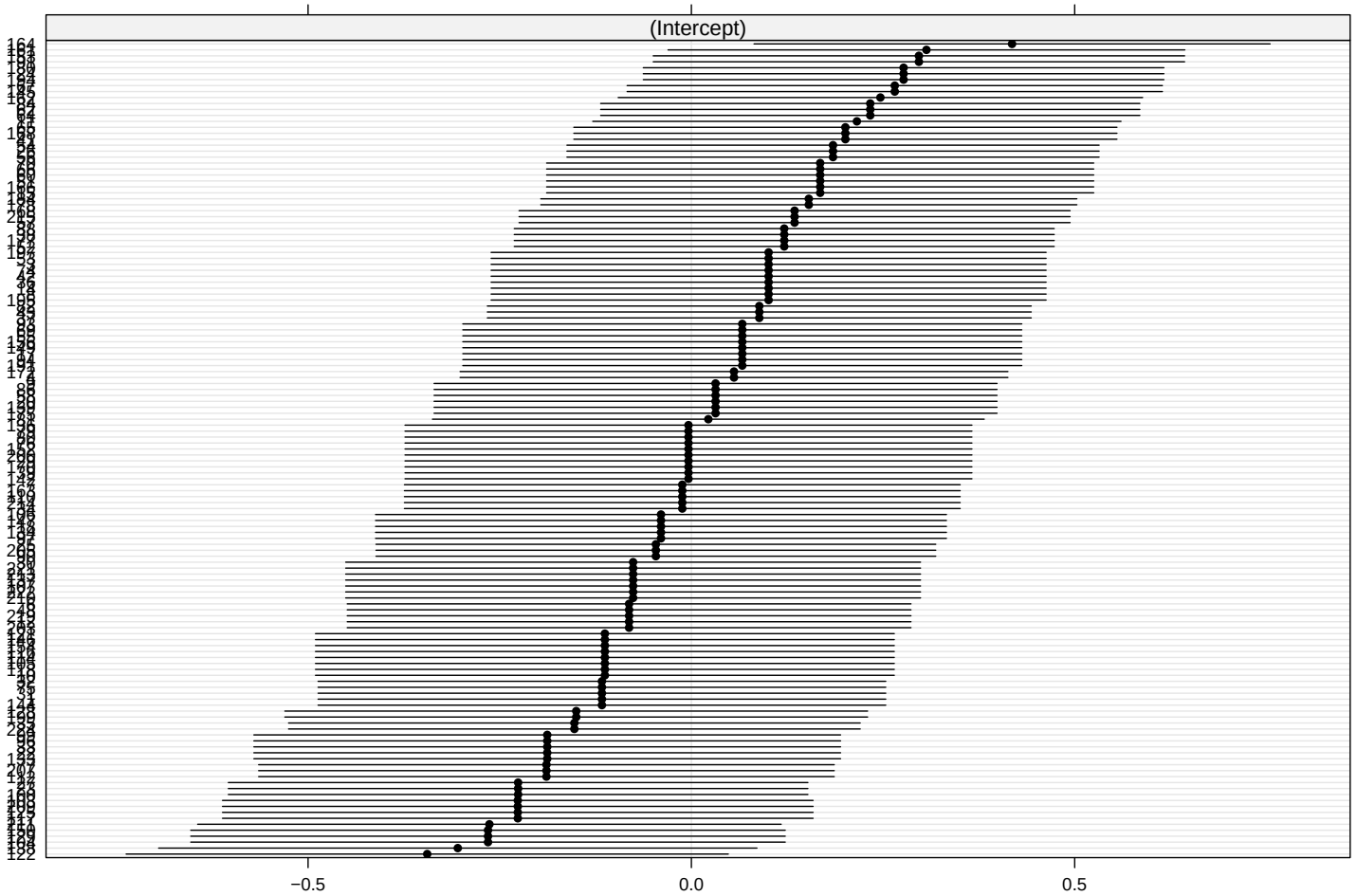
`grande` and `piccolo` are two nested models to be compared in order to test the significance of the three-way interaction.

```
>  ### "big" model with three-way interactions
> grande <- glmer( iz ~ type*time*quad + (1|pat) , data=dati14, family=poisson)
> ### "small" model with two-way interactions only
> piccolo <- glmer( iz ~  type*time + type*quad + time*quad + (1|pat) , data=dati14, family=poisson)
> ### comparison of the two models
> anova(piccolo, grande)
Data: dati14
Models:
piccolo: iz ~ type * time + type * quad + time * quad + (1 | pat)
grande: iz ~ type * time * quad + (1 | pat)
        Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
piccolo 14 1312.5 1382.5 -642.23
grande  17 1308.7 1393.7 -637.34 9.7792      3    0.02054 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Result: mild significance of the three-way interaction, i.e. of the necessity of the "big" model `grande` (p-value ≈ 0.02).

Random effects and their prediction intervals

Patient effects in the Poisson regression model are shown in the following graph



Esempio IZ: modello finale

```
> summary(grande)
Generalized linear mixed model fit by the Laplace approximation
Formula: iz ~ type * time * quad + (1 | pat)
   Data: dati14
  AIC  BIC logLik deviance
 1309 1394 -637.3     1275
Random effects:
 Groups Name        Variance Std.Dev.
 pat    (Intercept) 0.058711 0.24230
Number of obs: 1096, groups: pat, 137

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.81849    0.10127   8.083 6.34e-16 ***
type4                -0.21853    0.13029  -1.677   0.0935 .
timebef              -0.08377    0.13717  -0.611   0.5414
quaddr               -0.21869    0.14224  -1.537   0.1242
quadvl               -0.71118    0.16544  -4.299 1.72e-05 ***
quadvr               -0.44184    0.15178  -2.911   0.0036 **
type4:timebef         0.09568    0.17563   0.545   0.5859
type4:quaddr          0.21268    0.17992   1.182   0.2372
type4:quadvl          0.14706    0.20990   0.701   0.4835
type4:quadvr         -0.50176    0.21127  -2.375   0.0175 *
timebef:quaddr       -0.08532    0.20813  -0.410   0.6818
timebef:quadvl       -0.03174    0.24037  -0.132   0.8949
timebef:quadvr       -0.22260    0.22779  -0.977   0.3285
type4:timebef:quaddr -0.14823    0.26275  -0.564   0.5727
type4:timebef:quadvl  0.16643    0.29939   0.556   0.5783
type4:timebef:quadvr  0.76369    0.29734   2.568   0.0102 *
```

Da risposta/predittori a tante risposte correlate

La presenza di effetti casuali è uno spunto per diverse generalizzazioni:

–– non più una risposta tanti predittori ma tante variabili intercorrelate

–– collegamento con i modelli gerarchici bayesiani

Esempio: gene G3BP2

Interest on gene expression and gene regulation of gene G3BP2 and its genomic neighborhood.

differential expression $\longrightarrow$ inference on constrasts between line/gene interactions in BHM

gene regulation $\longrightarrow$ inference on regression coefficients between gene expressions

II database Melanoma

September HPRT reference gene

November HPRT reference gene

8 gene expression variables, in replicates of 3

line	Cdkl21a	Cdkl21b	G3BP21a	G3BP21b	G3BP21c	G3BP2as	C13sv	C13uns	HPRT-Sep	HPRT-Nov
MelA	25.95	26.26	23.46	26.01	29.26	26.93	34.52	27.42	20.73	21.77
MelA	25.96	26.27	23.36	25.83	30.06	27.03	34.15	27.91	20.79	21.88
MelA	NA	26.40	23.39	26.11	29.47	27.27	34.33	28.19	20.79	21.96
MelB	25.25	25.60	23.34	25.83	30.91	27.07	29.76	26.95	20.60	21.55
MelB	25.14	25.72	23.48	25.83	29.50	26.89	29.25	26.82	20.58	21.56
MelB	NA	25.59	23.46	26.04	29.77	27.06	29.27	26.92	20.62	21.47
MelC	25.12	25.39	22.82	25.71	29.52	27.14	34.61	24.39	21.06	21.48
MelC	24.58	25.23	22.79	25.63	30.11	27.22	34.02	24.67	21.08	21.40
MelC	NA	25.34	22.77	25.62	29.49	26.71	34.20	24.22	21.09	21.31
Skin	28.99	29.33	28.23	31.31	35.54	27.95	39.05	28.77	24.24	26.07
Skin	27.92	29.48	28.48	31.18	33.78	27.65	37.70	22.53	24.22	26.00
Skin	NA	29.11	28.30	31.31	33.87	26.97	36.81	22.48	24.19	26.16

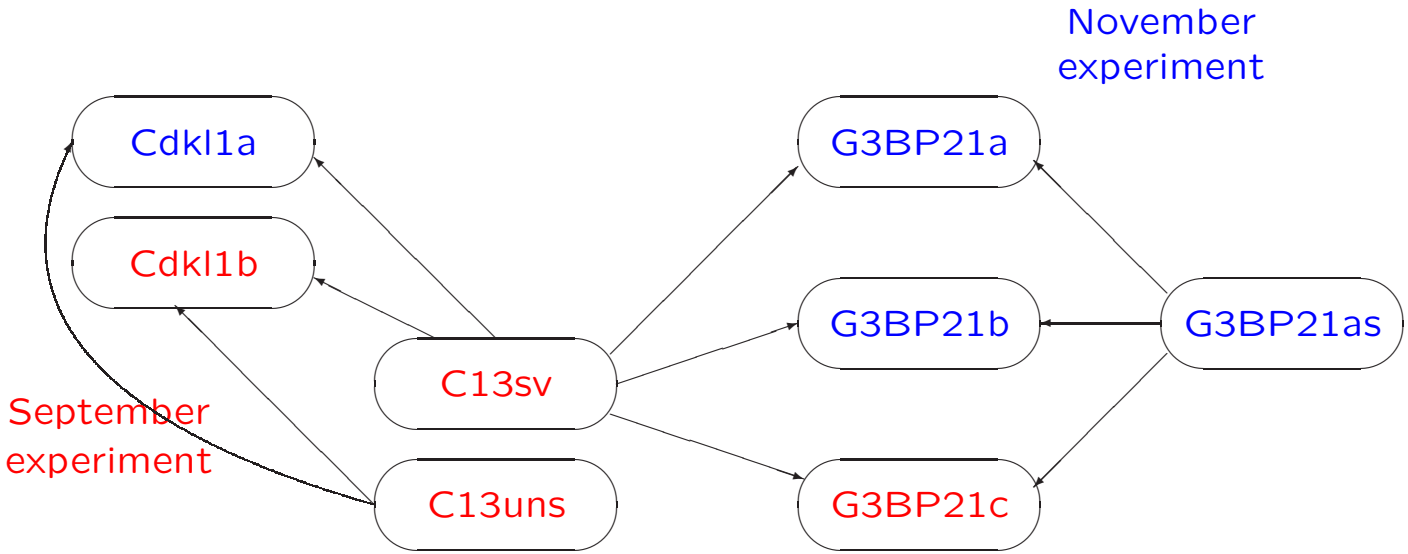
3 melanoma cell lines

one healthy skin cell line

Missing values

Gene regulation graph

The following quasi causal graph indicates the regression relationships of interest. Not yet DAG for the problem (the parameters are not yet there). Actually, we will avoid writing the full DAG and use instead its algebraic version.



Bayesian Hierarchical Models or, more in general, Statistical Bayesian Networks, address the same kinds of grouped data as mixed models (longitudinal data, repeated measurements, multielevel data, blocked designs, nested data) but from a Bayesian point of view.

Bayesian software: WinBUGS (OpenBUGS)

WinBUGS, just one of many (now) software for Bayesian computing (Lunn et al. (2000).

Declarative language which translates previous distributional assumptions into symbolic statements.

Then, based on MCMC, it computes posterior distributions.

It can be called within R using the `R2WinBUGS` library.

L'istruzione più utile in WinBUGS

`variabile ~ distribuzione`

Nell'impostazione bayesiana, sia le osservazioni che i parametri, i dati mancanti, i dati futuri, le caratteristiche latenti ecc... sono variabili aleatorie.

WinBUGS file for the melanoma case study

```
# modello.txt - by MG, March 2011
#               WinBUGS model file for the Australian project, mela part

########################## OBSERVATIONS
model;
{
### variable Cdkl21a
for ( i in 1:4 ) { ### tissue 1=MelA 2=MelB 3=MelC 4=Skin
for ( j in 1:3 ) { ### replicates 1,2,3
Cdkl21a[i,j] ~ dnorm( musuperCdkl21a[i], precsubrep ) } ### the Cdkl2-1a observations
musuperCdkl21a[i] <- csi + nusuperCdkl21a + tau[i] + gammasuperCdkl21a[i] + BetasubN * mus
}
                  [ ........ lines omitted ..........]

########################## PARAMETERS

### grand mean, random tissue effects
csi ~ dnorm( 25, 1.0E-3) ### informative prior to speed up simulations
nusuperN ~ dnorm( 22 , 1.0E-4)
nusuperS ~ dnorm( 22 , 1.0E-4)
nusuperCdkl21a ~ dnorm( 0 , precsubgene )
                  [ ........ lines omitted ..........]

for ( i in 1:3 ) { contsuperC13uns[i] <- gammasuperC13uns[i] -  gammasuperC13uns[4] }
}
```

R2WinBUGS file for the melanoma case study

```
# mela.R - by MG, March 2011
# R to WINBUGS code for the Australian project, mela part

################################### THE DATA ### read with source("mela.R")

dati <- list(
Cdkl21a = matrix( c(
25.95, 25.96, NA, 25.25, 25.14, NA, 25.12, 24.58, NA, 28.99, 27.92, NA
), c(4,3),byrow=T),                    [ ........ lines omitted ...........]

################################### THE INITIAL VALUES
inizio <- function() {list(
### initial missing values
Cdkl21a = matrix( c(
NA,NA,25,NA,NA,25,NA,NA,25,NA,NA,25), c(4,3),byrow=T),
### mean effects
nusuperCdkl21a =0,                     [ ........ lines omitted ...........]
)}
################################### CALLING WinBUGS
library(R2WinBUGS)
risult <- bugs(data=dati, inits=inizio, model.file=current,
n.chains=5,
n.iter=100000,                         [ ........ lines omitted ...........]
bugs.directory="c:/Programmi/WinBUGS14/"
)
```