GEE and lineal mixed models

Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology
Barcelona Institute for Global Health (ISGlobal)
e-mail:juanr.gonzalez@isglobal.org
http://brge.isglobal.org
and Departament of Mathematics, UAB

4 de junio de 2018

Modelos GEE y modelos lineales mixtos

- Datos longitudinales recogen observaciones repetidas de la variable respuesta a lo largo del tiempo, en un mismo individuo
- El analisis correcto de estos datos contempla que la correlacion entre las medidas de cada sujeto es tenida en cuenta
- A parte de las aproximaciones tradicionales (ANOVA para medidas repetidas, MANOVA, ...), tambien se puede:
  - Utilizar *Ecuaciones de Estimacion Generalizadas*: GEE
  - Modelos lineales mixtos
  - Modelos no-parametricos

Modelos GEE y modelos lineales mixtos

GEE

- Modelan la esperanza marginal o poblacional incorporando la correlacion entre las observaciones correspondientes a un mismo individuo, y se asume independencia de los individuos
- Admiten que la variable respuesta siga una distribucion distinta a la Gausiana
- Consideran una ecuacion de estimacion que se escribe en dos partes: una para modelar los parametros de regresion y la segunda para modelar la correlacion
- son bastante flexibles ya que el modelo solo necesita explicitar una funcion "link", una funcion de varianza y una estructura de correlacion

GEE

- Funcionan bien cuando:
  - el numero de observaciones por sujeto es pequeno y el numero de sujetos es grande
  - se tratan estdios longitudinales donde las medidas siempre se toman en el mismo instante de tiempo para todos los sujetos

GEE: Formulacion

1. Parte sistematica [lo mismo que un GLM]

$$g(E(Y_{ij})) = g(\mu_{ij}) = \beta' X_{ij}$$

   donde $i = 1, \ldots, n$ y $j = 1, \ldots, n_i$, y n denota el numero de individuos, y n_i el numero de medidas repetidas para el individuo i-esimo

2. Parte aleatoria

$$V(Y_{ij}) = \nu(\mu_{ij})\phi$$

   donde ν es la funcion de la varianza y ϕ el parametro de escala

3. Ademas se tiene que explicitar la estructura de la correlacion mediante la *working correlation matrix*, $R(\alpha)$

Modelos GEE y modelos lineales mixtos

GEE

- No es necesaria la especificacion de un modelo estadistico. Es decir, no es necesario conocer $f(y|parametros)$. Asi, son flexibles, pero:
  - la estimacion de las β's no tiene porque se la mejor posible
  - la inferencia esta basada en resultados asintoticos
  - los metodos de validacion son complicados
- La estimacion de los parametros se puede encontrar en muchos sitios (ver por ejemplo Liang y Zeger, Biometrika, 1986 o Zeger et al, Biometrics, 1988)
- si hay datos faltantes (missing) la estimacion solo es correcta si los missing son MCAR (missing completely at Random)

GEE con R

Para realizar todos los analisis se necesitan los datos en formato largo.

```
datos <- read.table("../../data/hypothetical_largo.txt", header=TRUE)
datos[1:12,]

##    id time score group
## 1   1    1    31     A
## 2   1    2    29     A
## 3   1    3    15     A
## 4   1    4    26     A
## 5   2    1    24     A
## 6   2    2    28     A
## 7   2    3    20     A
## 8   2    4    32     A
## 9   3    1    14     A
## 10  3    2    20     A
## 11  3    3    28     A
## 12  3    4    30     A
```

GEE con R

Cargamos la libreria

```
library(gee)
```

Usaremos la funcion `gee`

```
args(gee)

## function (formula = formula(data), id = id, data = parent.frame(),
##     subset, na.action, R = NULL, b = NULL, tol = 0.001, maxiter = 2
##     family = gaussian, corstr = "independence", Mv = 1, silent = TR
##     contrasts = NULL, scale.fix = FALSE, scale.value = 1, v4.4compa
## NULL
```

GEE con R

Antes de estimar el modelo:

- La funcion `gee` **asume** que los datos estan ordenados segun el individuo
- La esctructura de correlacion puede ser: independence, fixed, stat_M_dep, non_stat_M_dep, exchangeable, AR-M and unstructured

independence
: Es la eleccion mas sencilla e ineficiente, ignorando las medidas repetidas.

exchangeable
: es la tambien llamada estructura de simetria compuesta o esferica, o estructura de efectos aleatorios $Cov(X_{il}, Y_{ik}) = \alpha$. En este caso todas las correlaciones se suponen iguales:

AR-M
: de orden uno (M=1): $Cov(X_{il}, Y_{ik}) = \alpha^{|l-k|}$

unestructured
: Todas las correlaciones pueden ser diferentes. Adecuada si hay datos suficientes para estimar todas las varianzas-covarianzas

GEE con R

El modelo que asume independencia se puede estimar mediante:

```
mod.gee.indep <- gee(score ~ group + time,
                     data = datos, id = id,
                     family = gaussian,
                     corstr = "independence")

## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

Un modelo autoregresivo

```
mod.gee.AR <- gee(score ~ group + time,
                  data = datos, id = id,
                  family = gaussian,
                  corstr = "AR-M")

## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

GEE con R

Guardamos el summary (es largo)

```
ss.indep <- summary(mod.gee.indep)
ss.AR <- summary(mod.gee.AR)
names(ss.AR)

##  [1] "call"                "version"             "nobs"
##  [4] "residual.summary"    "model"               "title"
##  [7] "coefficients"        "working.correlation" "scale"
## [10] "error"               "iterations"
```

GEE con R

...y comparamos. Por ejemplo los efectos de las variables

```
ss.indep$coef

##              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept) 23.2916667   3.258980 7.1469197    3.265145 7.1334259
## groupB       4.5833333   2.463557 1.8604534    2.042375 2.2441192
## time         0.5833333   1.101736 0.5294673    1.099095 0.5307398

ss.AR$coef

##              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept) 23.3112357   3.245726 7.1821338    3.266573 7.1362980
## groupB       4.5786421   2.444581 1.8729759    2.041405 2.2428880
## time         0.5726056   1.098854 0.5210936    1.101360 0.5199076
```

GEE con R

O la *working correlation matrix*

```
ss.indep$working.correlation

##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1

ss.AR$working.correlation

##               [,1]          [,2]          [,3]          [,4]
## [1,]  1.000000e+00 -0.0102881605  0.0001058462 -1.088963e-06
## [2,] -1.028816e-02  1.0000000000 -0.0102881605  1.058462e-04
## [3,]  1.058462e-04 -0.0102881605  1.0000000000 -1.028816e-02
## [4,] -1.088963e-06  0.0001058462 -0.0102881605  1.000000e+00
```

Modelos lineales mixtos Podriamos usar un modelo lineal, pero:

- Las observaciones repetidas en cada grupo o cluster, no son necesariamente independientes.
- Con frecuencia, no solo se quieren tomar decisiones respecto de los grupos o cluster observados, sino que se quiere valorar el efecto de las variables explicativas en una poblacion de la que los grupos son una muestra.
- Puede ser de interes valorar la variacion del efecto de x de un grupo a otro.
- La estimacion del efecto medio de las variables explicativas en cada grupo puede ser muy deficiente si no se recoge la posible variabilidad entre los grupos.

Modelos GEE y modelos lineales mixtos

Modelos lineales mixtos

- Modeliza la realacion entre la variable dependiente y las covariables
- Estima la correlacion intra-individuo (se puede especificar una estructura)
- Se pueden aplicar a muchas situaciones (datos multinivel, ANOVA, datos longitudinales)
- No requieren puntos equidistantes (son covariables - se modeliza el efecto)
- Son robustos ante los missing

Modelos lineales mixtos

Un modelo mixto se puede representar como:

$$y = X\beta + Zu + \epsilon$$

donde

- y son las observaciones, con media $E(y) = X\beta$
- β es un vector de efectos fijos
- u is un vector i.i.d de variables aleatorias con media $E(u) = 0$ y matriz de varianza-covarianza $\text{var}(u) = G$
- ϵ es un vector de terminos i.i.d. correspondientes al error aleatorio con media $E(\epsilon) = 0$ y varianza $\text{var}(\epsilon) = R$
- X and Z son matrices de regresores que relacionan las observaciones y con β y u

Modelos lineales mixtos con R

- Modelo sencillo para interpretar (modelo lineal mixto con intercept aleatorio)

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + a_{ij} + \epsilon_{ij}$$

$$a_i \; N(0, \tau_a^2) \; , \tau_a^2 \geq 0$$

$$\epsilon_{ij} \; N(0, \tau^2) \; , \tau^2 > 0$$

- El modelo presenta ahora un intercept aleatorio (centrado en 0) que depende del individuo *i*-esimo
- La varianza del efecto aleatorio recoge la variabilidad entre los diferentes individuos
- La varianza del error recoge la variabilidad dentro de cada individuo no explicada por el modelo. NOTA: si la varianza del efecto aleatorio fuese nula, el modelo coincidiria con el modelo de efectos fijos o de regresion lineal.

Modelos lineales mixtos con R

Necesitamos la libreria `nlme`

```
library(nlme)
```

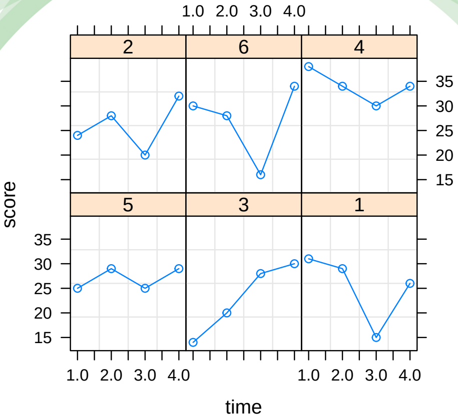
Debemos especificar la estructura de los datos mediante la funcion `groupedData`

```
datos.s <- groupedData(score ~ time | id, datos)
head(datos.s)

## Grouped Data: score ~ time | id
##   id time score group
## 1  1    1    31     A
## 2  1    2    29     A
## 3  1    3    15     A
## 4  1    4    26     A
## 5  2    1    24     A
## 6  2    2    28     A
```

Modelos lineales mixtos con R

Usa la libreria `trellis` para graficar (muy potente)

```
plot(datos.s)
```



Modelos lineales mixtos con R

El modelo de intercept aleatorio puede estimarse con:

```
mod.lme <- lme(score ~ time + group, datos.s, random =  ~ 1)
mod.lme

## Linear mixed-effects model fit by REML
##   Data: datos.s
##   Log-restricted-likelihood: -71.72926
##   Fixed: score ~ time + group
## (Intercept)        time      groupB
##  23.2916667   0.5833333   4.5833333
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:   0.5899484 6.012446
##
## Number of Observations: 24
## Number of Groups: 6
```

Modelos lineales mixtos con R

Comparamos con un modelo lineal

```
mod.lm <- lm(score ~ time + group, datos)
summary(mod.lm)

##
## Call:
## lm(formula = score ~ time + group, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.625  -3.708   0.375   3.938   9.542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.2917     3.2590   7.147 4.78e-07 ***
## time          0.5833     1.1017   0.529   0.6020
## groupB        4.5833     2.4636   1.860   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.034 on 21 degrees of freedom
## Multiple R-squared:  0.1512,Adjusted R-squared:  0.07039
## F-statistic: 1.871 on 2 and 21 DF,  p-value: 0.1788
```

Modelos lineales mixtos con R

El modelo con intercept y pendiente aleatoria puede estimarse con:

```
mod.lme2 <- lme(score ~ time + group, datos.s)
```

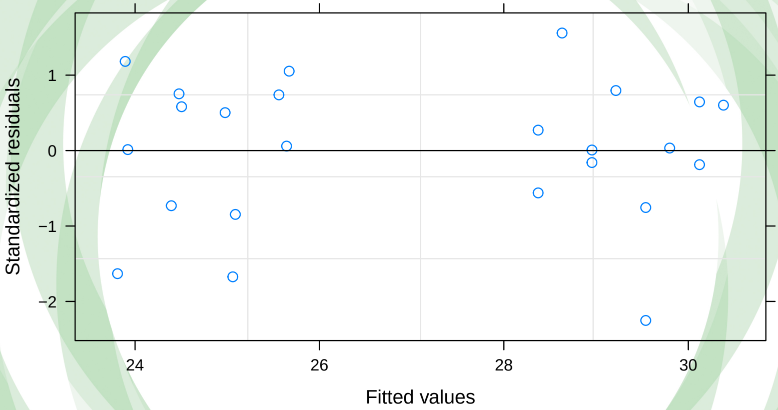
Â¿que metodo es el correcto?

```
anova(mod.lme, mod.lme2)

##           Model df      AIC      BIC    logLik   Test  L.Ratio p-val
## mod.lme       1  5 153.4585 158.6811 -71.72926
## mod.lme2      2 10 161.6750 172.1203 -70.83752 1 vs 2 1.783475  0.87
```

Modelos lineales mixtos con R

Model checking

```r
plot(mod.lme)
```

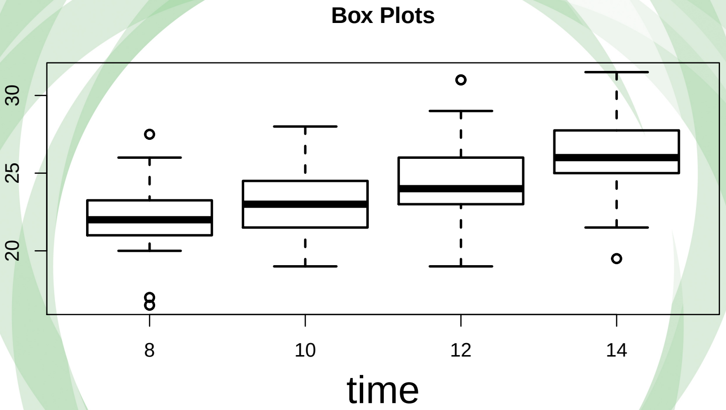


Orthodont: changes in orthodontic measurement over time

```
library(nparLD)

## Loading required package: MASS

data(Orthodont, package="nlme")
head(Orthodont)

## Grouped Data: distance ~ age | Subject
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

Metodo no parametrico para datos longitudinales

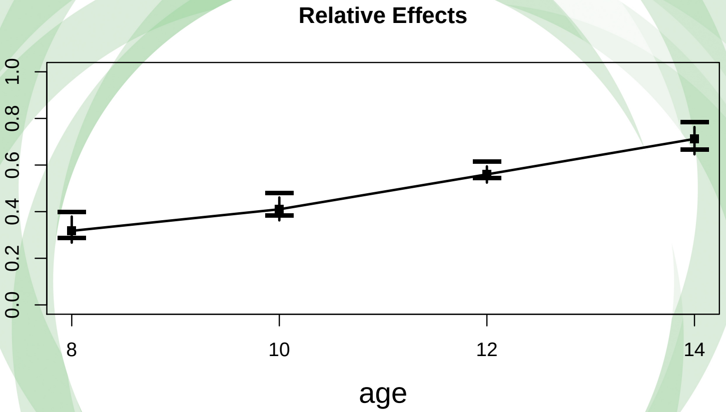
```
boxplot(distance ~ age, data = Orthodont, lwd = 2, xlab = "time",
    font.lab = 2, cex.lab = 2, main = "Box Plots")
```



Metodo no parametrico para datos longitudinales

```
mod0 <- nparLD(distance ~ age, data = Orthodont,
                 subject = "Subject", description = FALSE)

##  LD F1 Model
##  -----------------------
##  Check that the order of the time level is correct.
##  Time level:   8 10 12 14
##  If the order is not correct, specify the correct order in time.ord
```

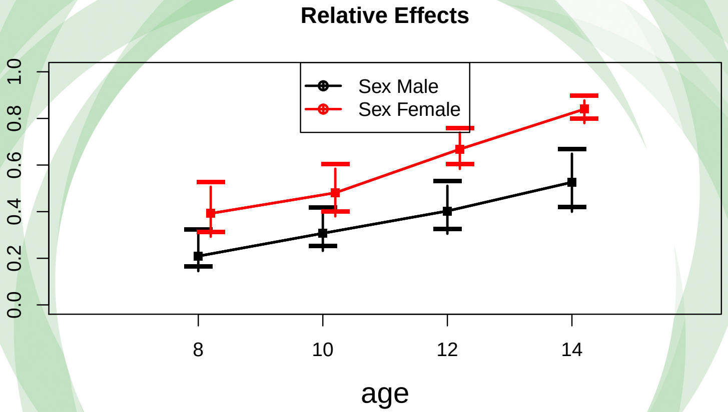
```r
plot(mod0)
```



Metodo no parametrico para datos longitudinales

```
mod1 <- nparLD(distance ~ age*Sex , data = Orthodont,
                 subject = "Subject", description = FALSE)

##  F1 LD F1 Model
##  -----------------------
##  Check that the order of the time and group levels are correct.
##  Time level:   8 10 12 14
##  Group level:   Male Female
##  If the order is not correct, specify the correct order in time.ord
```

```r
plot(mod1)
```


Relative Effects
Sex Male
Sex Female
0.0
0.2
0.4
0.6
0.8
1.0
8
10
12
14
age

Metodo no parametrico para datos longitudinales

```
summary(mod1)

## Model:
## F1 LD F1 Model
##
## Call:
## distance ~ age * Sex
##
## Relative Treatment Effect (RTE):
##                     RankMeans Nobs       RTE
## SexMale              64.79688   64 0.5953414
## SexFemale            39.52273   44 0.3613215
## age8                 32.98295   27 0.3007681
## age10                43.05966   27 0.3940709
## age12                58.32812   27 0.5354456
## age14                74.26847   27 0.6830414
## SexMale:age8         42.87500   16 0.3923611
## SexMale:age10        52.43750   16 0.4809028
## SexMale:age12        72.65625   16 0.6681134
## SexMale:age14        91.21875   16 0.8399884
## SexFemale:age8       23.09091   11 0.2091751
## SexFemale:age10      33.68182   11 0.3072391
## SexFemale:age12      44.00000   11 0.4027778
## SexFemale:age14      57.31818   11 0.5260943
```

```
Wald-Type Statistc (WTS):
         Statistic df       p-value
Sex       8.797738  1 3.016043e-03
age     103.424543  3 2.851266e-22
Sex:age   4.676974  3 1.970375e-01

ANOVA-Type Statistc (ATS):
        Statistic      df      p-value
Sex      8.797738 1.00000 3.016043e-03
age     46.191394 2.55914 7.475954e-26
Sex:age  1.872467 2.55914 1.412992e-01

Modified ANOVA-Type Statistic for the Whole-Plot Factors
    Statistic df1      df2     p-value
Sex  8.797738   1 17.57258 0.008431029
```