Análisis de datos longitudinales continuos (II)

Juan Ramón González
(juanr.gonzalez@isglobal.org)

Departamento de Matemáticas, Universidad Autónoma de Barcelona (UAB)
Insituto de Salud Global Barcelona (ISGlobal)

11 de mayo de 2017

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 análisis correcto de estos datos contempla que la correlación entre las medidas de cada sujeto es tenida en cuenta
- A parte de las aproximaciones tradicionales (vistas en la clase anterior), también se puede:
  - Utilizar *Ecuaciones de Estimación Generalizadas*: GEE
  - Modelos lineales mixtos

Modelos GEE y modelos lineales mixtos

GEE

- Modelan la esperanza marginal o poblacional incorporando la correlación entre las observaciones correspondientes a un mismo individuo, y se asume independencia de los individuos
- Admiten que la variable respuesta siga una distribución distinta a la Gausiana
- Consideran una ecuación de estimación que se escribe en dos partes: una para modelar los parametros de regresión y la segunda para modelar la correlación
- son bastante flexibles ya que el modelo sólo necesita explicitar una función "link", una función de varianza y una estructura de correlación

Modelos GEE y modelos lineales mixtos

GEE

- Funcionan bien cuando:
  - el número de observaciones por sujeto es pequeño y el número 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: Formulación

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 número de individuos, y n_i el número de medidas repetidas para el individuo i-ésimo

2. Parte aleatoria

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

   donde ν es la función de la varianza y ϕ el parámetro de escala

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

GEE

- No es necesaria la especificación de un modelo estadístico. Es decir, no es necesario conocer $f(y|parmetros)$. Así, son flexibles, pero:
    - la estimación de las β's no tiene porqué se la mejor posible
    - la inferencia está basada en resultados asintóticos
    - los métodos de validación son complicados
- La estimación de los parámetros 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 estimación sólo es correcta si los missing son MCAR (missing completely at Random)

GEE con R

Para realizar todos los análisis se necesitan los datos en formato largo. Usaremos los del seminario anterior

```
> datos <- read.table("../data/hypothetical_largo.txt",
> 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 librería

```
> library(gee)
```

Usaremos la función `gee`

```
> args(gee)
```

```
function (formula = formula(data), id = id, data = paren
    subset, na.action, R = NULL, b = NULL, tol = 0.001,
    family = gaussian, corstr = "independence", Mv = 1,
    contrasts = NULL, scale.fix = FALSE, scale.value = 1
NULL
```

GEE con R

Antes de estimar el modelo:

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

independence: Es la elección más sencilla e ineficiente, ignorando las medidas repetidas.

exchangeable: es la también llamada estructura de simetría compuesta o esférica, 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 la instrucción:

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

Un modelo autoregresivo

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

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 Como vimos en la sesión anterior, se podría 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 población de la que los grupos son una muestra.
- Puede ser de interés valorar la variación del efecto de x de un grupo a otro.
- La estimación 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 lineales mixtos

- Modeliza la realación entre la variable dependiente y las covariables
- Estima la correlación 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 $\mathrm{var}(u) = G$
- ϵ es un vector de términos i.i.d. correspondientes al error aleatorio con media $E(\epsilon) = 0$ y varianza $\mathrm{var}(\epsilon) = R$
- 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*-ésimo
- 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 coincidiría con el modelo de efectos fijos o de regresión lineal.

Modelos lineales mixtos con R

Necesitamos la librería `nlme`

```
> library(nlme)
```

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

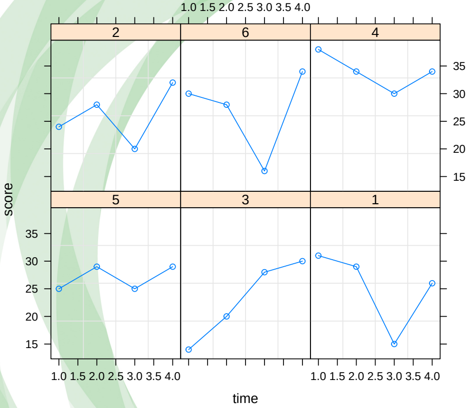
```
> datos.s <- groupedData(score ~ time | id, datos)
> head(datos.s)

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 librería `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)
```

¿cuál es necesario?

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

```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
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  0.8782
```

Modelos lineales mixtos con R

Model checking

```
> plot(mod.lme)
```