

MAT02035 - Modelos para dados correlacionados

Equações de Estimação Generalizadas - Exemplos

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Porto Alegre, 2022

Equações de estimação generalizadas

- ▶ Modelo de média populacional ou marginal, fornece uma abordagem de regressão para modelos lineares generalizados quando as respostas não são independentes (dados correlacionados/agrupados).
- ▶ O objetivo é fazer inferências sobre a população, levando em consideração a correlação das medidas dentro de indivíduo.
- ▶ Os pacotes gee e geepack são usados para modelos GEE no R.
- ▶ A principal diferença entre gee e geepack é que o geepack contém um método ANOVA que nos permite comparar modelos e realizar testes de Wald.

Equações de estimação generalizadas

- ▶ Sintaxe básica para `geeglm()` do pacote `geepack` (**tem uma sintaxe muito parecida com `glm()`**).

```
geeglm(formula, family = gaussian, data, id,  
       zcor = NULL, constr, std.err = "san.se")
```

- ▶ `formula` Descrição simbólica do modelo a ser ajustado;
- ▶ `family` Descrição da distribuição da resposta e função de ligação;
- ▶ `data` dataframe opcional;
- ▶ `id` vetor que identifica os *clusters* (agrupamentos);
- ▶ `zcor` especifica uma estrutura de correlação definida pelo usuário;
- ▶ `constr` estrutura de **correlação de trabalho**: “independence”, “exchangeable”, “ar1”, “unstructured”, “userdefined”;
- ▶ `std.err` tipo de erro padrão a ser calculado; o padrão “san.se” é a **estimativa robusta (sanduíche)**; use “jack” para obter uma estimativa da variância aproximada por *jackknife*.

Estrutura de correção

- *Independence* (independência),

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- *Exchangeable* (simetria composta),

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

Estrutura de correção

- ▶ *Autoregressive order 1* (autorregressivo de ordem 1),

$$\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

- ▶ *Unstructured* (não estruturada),

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

- ▶ O modelo GEE fornecerá resultados válidos com uma estrutura de correlação mal especificada quando o estimador de variância sanduíche for usado.

Inferência

- ▶ Para um objeto `geeglm` retornado por `geeglm()`, as funções `drop1()`, `confint()` e `step()` não se aplicam; no entanto `anova()` se aplica.
- ▶ A função `esticon()` no pacote `doBy` calcula e testa funções lineares dos parâmetros de regressão para objetos `lm`, `glm` e `geeglm`.
- ▶ Sintaxe básica:

```
esticon(obj, cm, beta0, joint.test = FALSE)
```

- ▶ `obj`: objeto do modelo;
- ▶ `L`: matriz especificando funções lineares dos parâmetros de regressão (uma função linear por linha e uma coluna para cada parâmetro);
- ▶ `beta0`: vetor de números (H_0 para β);
- ▶ `joint.test`: se `TRUE` um teste de hipóteses de Wald conjunto $L\beta = \beta_0$ é realizado, *default* é um teste para cada linha, $(L\beta)_i = \beta_{0,i}$.

Exemplo - GEE

```
# Instala e carrega os pacotes geepack e doBy
install.packages("geepack")
install.packages("doBy")
library(geepack)
library(doBy)
# conjunto de dados ohio do geepack - Efeito da poluição do ar na saúde
# Crianças acompanhadas por quatro anos, com chiado registrado anualmente
data(ohio) # carrega o conjunto de dados
head(ohio)
str(ohio)
# Variável resposta é binária - ajuste um modelo GEE logístico
# tempo (idade; age) como var. contínua
fit.exch <- geeglm(resp ~ age + smoke,
                  family = binomial(link = "logit"),
                  data = ohio, id = id,
                  corstr = "exchangeable", std.err = "san.se")
fit.unstr <- geeglm(resp ~ age + smoke,
                  family = binomial(link = "logit"),
                  data = ohio, id = id,
                  corstr = "unstructured", std.err = "san.se")

summary(fit.exch)
summary(fit.unstr)
```

Exemplo - GEE

```
# tempo (idade; age) como var. categórica
fit <- geeglm(resp ~ factor(age) + smoke,
              family = binomial(link = "logit"),
              data = ohio, id = id,
              corstr = "exchangeable", std.err = "san.se")

summary(fit)
# Teste o efeito de smoke usando anova()
fit1 <- geeglm(resp ~ factor(age) + smoke,
               family = binomial(link = "logit"),
               data = ohio, id = id,
               corstr = "exchangeable", std.err = "san.se")

fit2 <- geeglm(resp ~ factor(age),
               family = binomial(link = "logit"),
               data = ohio, id = id,
               corstr = "exchangeable", std.err = "san.se")

anova(fit1, fit2)
# Teste Wald individual e intervalo de confiança para cada parâmetro
est <- esticon(fit, diag(5))
# Odds ratio and confidence intervals
OR.CI <- exp(cbind(est$estimate, est$lwr, est$upr))
rownames(OR.CI) <- names(coef(fit))
colnames(OR.CI) <- c("OR", "OR 95% LI", "OR 95% LS")
```


Exemplo - GEE

```

# Razão de chances de chiado no peito para uma criança de 9 anos com uma mãe que
# fumou durante o primeiro ano do estudo em comparação com uma criança de 8
# anos com uma mãe que não fumou durante o primeiro ano do estudo.
# Isto é, estimar [smoke+factor(age)0] - [factor(age)-1]
esticon(fit, c(0,-1,1,0,1))
exp(.Last.value$estimate)
# 9 anos de idade com mãe que fumava tem maior risco de chiado no peito
# Teste conjuntamente os efeitos usando esticon()
fit <- geeglm(resp ~ factor(age)*smoke,
              family = binomial(link = "logit"),
              data = ohio, id = id,
              corstr = "exchangeable", std.err = "san.se")

summary(fit)
L <- cbind(matrix(0, nrow=3, ncol=5), diag(3))
esticon(fit, L, joint.test=TRUE)
# Também poderia usar anova()
fit1 <- geeglm(resp ~ factor(age)*smoke,
               family = binomial(link = "logit"),
               data = ohio, id = id,
               corstr = "exchangeable", std.err = "san.se")
fit2 <- geeglm(resp ~ factor(age) + smoke,
               family = binomial(link = "logit"),
               data = ohio, id = id,
               corstr = "exchangeable", std.err = "san.se")
anova(fit1, fit2)

```

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ Ensaio clínico controlado por placebo de 30 pacientes com hanseníase no Eversley Childs Sanitorium, nas Filipinas.
- ▶ Os participantes foram aleatorizados para um dos dois antibióticos (medicamento indicado para tratamento A e B) ou para um placebo (medicamento indicado para tratamento C).
- ▶ Os dados da linha de base sobre o número de bacilos da hanseníase em 6 locais do corpo foram registrados.
- ▶ Após vários meses de tratamento, o número de bacilos foi registrado pela segunda vez.
- ▶ Desfecho/resposta: contagem total do número de bacilos da hanseníase em 6 locais.

Ensaio Clínico de Antibióticos para Hanseníase

Table 1: Summary descriptives table by groups of '0=Drug C, 1=Drug A, 2=Drug B'

	0 N=10	1 N=10	2 N=10
y	12.9 (3.96)	9.30 (4.76)	10.0 (5.25)
y	12.3 (7.15)	5.30 (4.64)	6.10 (6.15)

Ensaio Clínico de Antibióticos para Hanseníase

Questão: O tratamento com antibióticos (medicamentos A e B) reduz a abundância de bacilos da hanseníase quando comparado ao placebo (medicamento C).

Consideramos o modelo a seguir para alterações na contagem média

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2 \text{tempo}_{ij} + \beta_3 \text{tempo}_{ij} \times \text{trt}_{1i} + \beta_4 \text{tempo}_{ij} \times \text{trt}_{2i},$$

em que Y_{ij} é a contagem de bacilos para o i -ésimo paciente no j -ésimo período ($j = 1, 2$).

- ▶ trt_1 e trt_2 são variáveis indicadoras para os medicamentos A e B, respectivamente.

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ A variável binária, tempo, denota os períodos de linha de base e pós-tratamento, com $\text{tempo} = 0$ para o período de linha de base (período 1) e $\text{tempo} = 1$ para o período de acompanhamento de pós-tratamento (período 2).

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ Para completar a especificação do modelo marginal, assumimos

$$\text{Var}(Y_{ij}) = \phi \mu_{ij},$$

em que ϕ pode ser pensado como um fator de super-dispersão.

- ▶ Por fim, a associação dentro de indivíduo é contabilizada assumindo uma correlação comum,

$$\text{Corr}(Y_{i1}, Y_{i2}) = \alpha.$$

- ▶ Os parâmetros de regressão log-linear, β , podem receber interpretações em termos de (log) razões de taxa.

Ensaio Clínico de Antibióticos para Hanseníase

Tratamento	Período	$\log(\mu_{ij})$
Droga A (antibiótico)	Linha de base	β_1
	Acompanhamento	$\beta_1 + \beta_2 + \beta_3$
Droga B (antibiótico)	Linha de base	β_1
	Acompanhamento	$\beta_1 + \beta_2 + \beta_4$
Droga C (placebo)	Linha de base	β_1
	Acompanhamento	$\beta_1 + \beta_2$

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ Por exemplo, e^{β_2} é a razão de taxas de bacilos da hanseníase, comparando o período de acompanhamento com a linha de base, no grupo placebo (medicamento C).
- ▶ Da mesma forma, $e^{\beta_2+\beta_3}$ é a razão de taxas correspondente no grupo aleatorizado para a droga A.
- ▶ Finalmente, $e^{\beta_2+\beta_4}$ é a razão de taxas correspondente no grupo aleatorizado para o medicamento B.
- ▶ Assim, β_3 e β_4 representam a diferença entre as mudanças nas log-taxas esperadas, comparando os medicamentos A e B com o placebo (medicamento C).

Ensaio Clínico de Antibióticos para Hanseníase

```
fit <- geeglm(y ~ tempo + tempoA + tempoB,
             family = poisson(link = "log"),
             data = ds.longo, id = id, waves = tempo,
             corstr = "exchangeable", std.err = "san.se")
summary(fit)
```

```
##
## Call:
## geeglm(formula = y ~ tempo + tempoA + tempoB, family = poisson(link = "log"),
##       data = ds.longo, id = id, waves = tempo, corstr = "exchangeable",
##       std.err = "san.se")
##
## Coefficients:
##             Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)  2.373354  0.080138 877.101  <2e-16 ***
## tempo        -0.002877  0.157005  0.000  0.9854
## tempoA       -0.562571  0.221983  6.423  0.0113 *
## tempoB       -0.495284  0.234201  4.472  0.0344 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   3.214  0.4998
## Link = identity
##
## Estimated Correlation Parameters:
##             Estimate Std.err
## alpha      0.7384  0.08149
## Number of clusters: 30 Maximum cluster size: 2
```

Ensaio Clínico de Antibióticos para Hanseníase

$H_0 : \beta_3 = \beta_4 = 0$ (as médias de A e B não diferem de C).

```
L <- rbind(c(0,0,1,0),
           c(0,0,0,1))
esticon(fit, L, joint.test = TRUE)
```

```
##    X2.stat DF Pr(>|X^2|)
## 1      7.34  2      0.02548
```

$H_0 : \beta_3 = \beta_4$ (as médias de A e B não diferem).

```
L <- c(0,0,1,-1)
esticon(fit, L, joint.test = TRUE)
```

```
##    X2.stat DF Pr(>|X^2|)
## 1 0.08848  1      0.7661
```

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ Para obter uma estimativa comum da log-razão de taxas, comparando ambos os antibióticos (medicamentos A e B) e o placebo, podemos usar o **modelo reduzido**

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2 \text{tempo}_{ij} + \beta_3 \text{tempo}_{ij} \times \text{trt}_i,$$

em que a variável trt é uma variável indicadora para uso de antibiótico, com $\text{trt} = 1$ se um paciente foi aleatorizado para o medicamento A ou B e $\text{trt} = 0$ caso contrário.

- ▶ Mantemos as mesmas suposições sobre a variância e correlação de antes.

Ensaio Clínico de Antibióticos para Hanseníase

```
fit <- geeglm(y ~ tempo + tempoAB,
             family = poisson(link = "log"),
             data = ds.longo, id = id, waves = tempo,
             corstr = "exchangeable", std.err = "san.se")

summary(fit)

##
## Call:
## geeglm(formula = y ~ tempo + tempoAB, family = poisson(link = "log"),
##       data = ds.longo, id = id, waves = tempo, corstr = "exchangeable",
##       std.err = "san.se")
##
## Coefficients:
##           Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  2.37335  0.08014 877.10  <2e-16 ***
## tempo       -0.00286  0.15700   0.00  0.9855
## tempoAB     -0.52783  0.19883   7.05  0.0079 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##           Estimate Std.err
## (Intercept)    3.23    0.52
## Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha      0.738    0.081
## Number of clusters: 30 Maximum cluster size: 2
```

Ensaio Clínico de Antibióticos para Hanseníase

	RT	RT 95% LI	RT 95% LS
(Intercept)	10.73	9.17	12.56
tempo	1.00	0.73	1.36
tempoAB	0.59	0.40	0.87

Ensaio Clínico de Antibióticos para Hanseníase

- ▶ A estimativa comum da log-taxa é de -0.52783.
- ▶ A razão de taxas é de 0.60 (ou $e^{-0.52783}$), com intervalo de confiança de 95%, 0.4 a 0.87, indicando que o tratamento com antibióticos reduz significativamente o número médio de bacilos quando comparado ao placebo.
- ▶ Para o grupo placebo, há uma redução não significativa no número médio de bacilos de aproximadamente 3% (ou $[1 - e^{-0.00286}] \times 100\%$).
- ▶ No grupo de antibióticos, há uma redução significativa de aproximadamente 41% (ou $[1 - e^{-0.00286 - 0.52783}] \times 100\%$).
- ▶ A correlação estimada em pares de 0.74 é relativamente grande.
- ▶ O parâmetro de escala estimado de 3.23 indica superdispersão substancial (se não há superdispersão, então $\phi = 1$).

Avisos

- ▶ **Próxima aula:** Modelos multiníveis.
- ▶ **Para casa (1):** ler o Capítulo 12 e 13 do livro “**Applied Longitudinal Analysis**”.
 - ▶ Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 e 11.
 - ▶ Veja o *help* do pacote *geepack* do R.
- ▶ **Para casa (2):** Faça o exercício 13.1 do Cap. 13 do livro “**Applied Longitudinal Analysis**”.

Bons estudos!

