Resolução da Lista 3 - Análise de Dados Longitudinais

Helen Lourenço e Vitor Kroeff

Questão 1

a)

```
dados <- tibble(foreign::read.dta("toenail.dta"))
ajuste_gee <-geeglm(y ~ trt * month - trt,family = binomial(link = "logit"),
                    data = dados,id = id,corstr = "exchangeable")
```

b)

```
exp(ajuste_gee$coefficients)
```

```
(Intercept)       month   trt:month
  0.5608934   0.8425524   0.9252438
```

O coeficiente β_2 está relacionado ao efeito do tempo no grupo de tratamento A (ou controle). O expoente e^{β_2} representa uma razão de chances. Essa razão de chances indica que, no grupo A, o coeficiente está associado a uma redução na probabilidade de ocorrência de onicólise com o passar dos meses.

c)

O coefiente β_3 está associado a interação entre o tratamento B e o tempo em meses. Assim como na alternativa anterior, vemos que $e^{\beta_3} = 0,925$ está assciado a uma reduzação das chances de ocorrência de onicólise com o passar dos meses, porém uma redução menor que a do grupo controle.

d)

Podemos observar o `summary` do modelo ajustado como sendo:

```r
summary(ajuste_gee)
```

```

Call:
geeglm(formula = y ~ trt * month - trt, family = binomial(link = "logit"),
    data = dados, id = id, corstr = "exchangeable")

 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -0.57822  0.13041 19.661 9.25e-06 ***
month       -0.17132  0.02957 33.574 6.86e-09 ***
trt:month   -0.07770  0.05379  2.086    0.149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    1.088  0.5265
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.4217  0.2203
Number of clusters:   294  Maximum cluster size: 7
```

Com base no p-valor associado ao β_3, não parece haver uma diferença significativa entre os tratamentos aplicados nos Grupos A e B. Também com base nos coeficientes do modelos, podemos observar que essa chance de desnvolver uma onicólise severa diminui com o passar dos meses.

e)

O modelo pode ser ajustado como:

```r
ajuste_misto <-lme4::glmer(
  formula = y ~ month + trt:month + (1 | id),
  family = binomial(link = "logit"),
  data = dados)
```

f)

Como o efeito aleatório está apenas no intercepto, vimos a seguinte relação aproximada entre entre os coeficientes do modelo marginal e aqueles do modelo de efeitos aleatórios.

$$\beta_M = \frac{\beta_{EA}}{\sqrt{1 + \frac{16\sqrt{3}}{15\pi}\sigma_b^2}}$$

```r
var_bi <- lme4::VarCorr(ajuste_misto)
var_bi
```

```
 Groups Name        Std.Dev.
 id     (Intercept) 4.54
```

```r
var_bi <- sqrt(as.numeric(var_bi)) # Conversão para númerico
fator <- sqrt(1 + (16*sqrt(3)/(15*pi)*var_bi))
fator
```

```
[1] 1.916
```

Podemos comparar as magnitudes dos efeitos da seguinte forma:

```r
b.gee <- summary(ajuste_gee)$coef[,1]
b.lme <- summary(ajuste_misto)$coef[,1]
p.gee <- summary(ajuste_gee)$coef[,4]
p.lme <- summary(ajuste_misto)$coef[,4]
round(cbind(b.gee, or.gee = exp(b.gee), p.gee, b.lme, or.lme = exp(b.lme), p.lme,
            razao.b = b.lme/b.gee), 3)
```

```
             b.gee or.gee p.gee  b.lme or.lme p.lme razao.b
(Intercept) -0.578  0.561 0.000 -2.649  0.071  0.00   4.581
month       -0.171  0.843 0.000 -0.396  0.673  0.00   2.309
month:trt   -0.078  0.925 0.149 -0.146  0.865  0.03   1.873
```

Podemos observar estimativas maiores para o modelo misto em relação ao GEE, mas com o mesmo sinal, indicando uma concordância dos efeitos das variáveis.

g)

```r
knitr::kable(
exp(summary(ajuste_misto)$coef)) # Exponencial dos parâmetros do modelo
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0707	2.031	0.0238	1.00
month	0.6733	1.047	0.0002	1.00
month:trt	0.8646	1.069	0.1134	1.03

A interpretação é muito próxima a do modelo GEE, onde β_2 está associada a variação no mês no grupo A (controle), porém está sendo levado em conta a variação de cada paciente do grupo A por conta do efeito aleatório no intercepto.

h)

O coeficiente β_3 está relacionado na interação entre o grupo B e o tempo em meses. Assim como comentado na alternativa anterior, está sendo levado em conta a variação dentro dos indivíduos do grupo por conta do efeito aleatório no intercepto do modelo ajustado.

i)

- Ajuste do modelo GEE

```r
summary(ajuste_gee)$coefficients
```

```
            Estimate Std.err   Wald  Pr(>|W|)
(Intercept)  -0.5782 0.13041 19.661 9.248e-06
month        -0.1713 0.02957 33.574 6.861e-09
trt:month    -0.0777 0.05379  2.086 1.486e-01
```

- Ajuste do modelo misto

```r
summary(ajuste_misto)$coef
```

```
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -2.6491    0.70868  -3.738 1.854e-04
month        -0.3956    0.04567  -8.663 4.606e-18
month:trt    -0.1455    0.06687  -2.176 2.952e-02
```

Podemos observar que o β_3 estimado no modelo GEE é consideravelmente menor em relação ao modelo misto. Além disso, a interação entre tratamento e tempo é significativa ao nível de 5% apenas no modelo misto.

Essa diferença nas estimativas ocorre devido à natureza dos modelos: o GEE utiliza uma matriz de correlação especificada (neste caso, simetria composta), focado estimativas marginais e ignorando a variação intra-indivíduo. Já o modelo misto incorpora efeitos aleatórios, que modelam a variação intra-indivíduo, resultando em estimativas condicionais aos indivíduos da base.

Questão 2

a)

Os modelos podem ser ajustados como:

```
ajuste_gee_ind <- geeglm(response ~ group * time,data = dados_rats,
                         id = subject,corstr = "independence")

ajuste_gee_simetria <- geeglm(response ~ group * time,data = dados_rats,
                         id = subject,corstr = "exchangeable")

ajuste_gee_ar1 <- geeglm(response ~ group * time,data = dados_rats,
                         id = subject,corstr = "ar1")

ajuste_gee_unstructured <- geeglm(response ~ group * time,data = dados_rats,
                         id = subject,corstr = "unstructured")
```

b)

Nos ajustes realizados na alternativa anterior Y_{ij}(`response`) segue uma distribuição Normal com diferentes formas de estimar a matriz de coreelção. Dentre as ajustas estão (em ordem):

- **Independete**:Nenhuma correlação entre as observações repetidas.

- **Simetria Composta**: Todas as observações dentro de um indivíduo têm a mesma correlação.
- **AR(1)**: As observações mais próximas no tempo têm maior correlação.
- **Não Estruturada**: A correlação entre cada par de observações é estimada de forma independente.

c)

Podemos comparar cada um dos modelos ajustados na alternativa **a)** e comparar com a sua respectiva estrutura do modelo `gls()` da seguinte forma:

Independente

Para os modelos com matriz de correlação independentes, podemos observar que o valor dos coeficientes é praticamento o mesmo, porém o p-valor associado as variáveis do modelo GLS aparenta ser menos significativo.

```r
dados_rats <- na.omit(dados_rats)

summary(ajuste_gee_ind)$coefficients # Ajuste GEE
```

```
             Estimate  Std.err     Wald Pr(>|W|)
(Intercept) 63.049929 1.186235 2825.059   0.0000
group        0.243826 0.678895    0.129   0.7195
time         0.203898 0.014009  211.857   0.0000
group:time  -0.008235 0.008066    1.042   0.3073
```

```r
ajuste_gls_ind <- gls(response ~ group * time,data = dados_rats)
summary(ajuste_gls_ind) # Ajuste GLS
```

```
Generalized least squares fit by REML
  Model: response ~ group * time
  Data: dados_rats
   AIC  BIC logLik
  1203 1221 -596.7

Coefficients:
            Value Std.Error t-value p-value
(Intercept) 63.05    1.6633   37.91  0.0000
```

```
group        0.24    0.8163    0.30  0.7654
time         0.20    0.0217    9.39  0.0000
group:time  -0.01    0.0108   -0.76  0.4476

 Correlation:
           (Intr) group  time
group      -0.924
time       -0.969  0.902
group:time  0.889 -0.969 -0.924

Standardized residuals:
     Min       Q1      Med       Q3      Max
-2.57774 -0.67881  0.04173  0.61316  2.60110

Residual standard error: 2.511
Degrees of freedom: 252 total; 248 residual
```

Simetria Composta

Novamente observamos um p-valor maior nos coeficientes do ajuste GLS. Nestes modelos podemos observar que o efeito de interação entre grupo e tempo é zero no modelo gls.

```r
summary(ajuste_gee_simetria)$coefficients # Ajuste GEE
```

```
             Estimate  Std.err      Wald Pr(>|W|)
(Intercept) 63.774279 1.135531 3.154e+03   0.0000
group       -0.326590 0.561568 3.382e-01   0.5609
time         0.190638 0.009604 3.940e+02   0.0000
group:time   0.001408 0.005473 6.618e-02   0.7970
```

```r
ajuste_gls_simetria <- gls(response ~ group * time,correlation= corCompSymm(form= ~1|subject)
                      data = dados_rats)
summary(ajuste_gls_simetria) # Ajuste GLS
```

```
Generalized least squares fit by REML
  Model: response ~ group * time
  Data: dados_rats
   AIC  BIC logLik
  1098 1119 -542.8
```

```
Correlation Structure: Compound symmetry
 Formula: ~1 | subject
 Parameter estimate(s):
   Rho
0.5768

Coefficients:
            Value Std.Error t-value p-value
(Intercept) 63.80    1.3541   47.11  0.0000
group       -0.34    0.6632   -0.52  0.6043
time         0.19    0.0156   12.23  0.0000
group:time   0.00    0.0078    0.22  0.8283

 Correlation:
           (Intr) group  time
group      -0.923
time       -0.818  0.771
group:time  0.749 -0.827 -0.924

Standardized residuals:
     Min       Q1      Med       Q3      Max
-2.50885 -0.67999  0.03162  0.64124  2.56961

Residual standard error: 2.569
Degrees of freedom: 252 total; 248 residual
```

AR(1)

No caso da estrutura de correlações AR(1), podemos observar que o erro padrão do estimador GLS é muito superior ao do GEE, indicando que os modelos GLS são mais sensíveis a estrutura de correlação.

```
summary(ajuste_gee_ar1)$coefficients # Ajuste GEE
```

```
             Estimate  Std.err      Wald Pr(>|W|)
(Intercept) 61.168528 1.095928 3.115e+03   0.0000
group        0.039572 0.601441 4.329e-03   0.9475
time         0.216039 0.012503 2.986e+02   0.0000
group:time  -0.001769 0.006971 6.441e-02   0.7997
```

```r
ajuste_gls_ar1 <- gls(response ~ group * time,correlation= corAR1(form= ~1|subject),
                      data = dados_rats)
summary(ajuste_gls_ar1) # Ajuste GLS
```

```
Generalized least squares fit by REML
  Model: response ~ group * time
  Data: dados_rats
   AIC  BIC logLik
  1090 1111 -538.8

Correlation Structure: AR(1)
 Formula: ~1 | subject
 Parameter estimate(s):
   Phi
0.6803

Coefficients:
            Value Std.Error t-value p-value
(Intercept) 61.41    1.9867   30.91  0.0000
group        0.07    0.9661    0.07  0.9440
time         0.21    0.0256    8.39  0.0000
group:time   0.00    0.0127   -0.22  0.8261

 Correlation:
           (Intr) group  time
group      -0.923
time       -0.934  0.871
group:time  0.854 -0.935 -0.923

Standardized residuals:
    Min      Q1     Med      Q3     Max
-2.1422 -0.4760  0.1491  0.7790  2.8588

Residual standard error: 2.575
Degrees of freedom: 252 total; 248 residual
```

Não Estruturada

O caso não estruturado parece ser o que apresenta maior diferença entre os dois modelos, tanto no valor dos coeficientes, quanto no erro padrão associado a eles. O efeito de grupo parece ser bem mais forte no estimador GEE em comparação ao GLS, mas com um erro padrão muito mais alto também.

```
summary(ajuste_gee_unstructured)$coefficients # Ajuste GEE
```

```
            Estimate Std.err    Wald Pr(>|W|)
(Intercept) 70.55828 2.35043 901.157  0.00000
group       -2.50386 1.05934   5.587  0.01810
time         0.08166 0.04398   3.448  0.06335
group:time   0.03421 0.01848   3.428  0.06410
```

```
ajuste_gls_unstructured <- gls(response ~ group * time,correlation= corSymm(form= ~1|subject)
                      data = dados_rats)
summary(ajuste_gls_unstructured) # Ajuste GLS
```

```
Generalized least squares fit by REML
  Model: response ~ group * time
  Data: dados_rats
   AIC  BIC logLik
  1032 1124 -490.2

Correlation Structure: General
 Formula: ~1 | subject
 Parameter estimate(s):
 Correlation:
  1     2     3     4     5     6
2 0.323
3 0.302 0.846
4 0.355 0.830 0.814
5 0.601 0.742 0.685 0.604
6 0.796 0.520 0.481 0.503 0.721
7 0.838 0.468 0.361 0.341 0.692 0.890

Coefficients:
            Value Std.Error t-value p-value
(Intercept) 62.76    0.9430   66.56  0.0000
group       -0.14    0.4592   -0.30  0.7610
time         0.20    0.0088   22.87  0.0000
group:time   0.00    0.0045   -0.06  0.9513

 Correlation:
```

```
           (Intr) group  time
group      -0.923
time       -0.592  0.569
group:time  0.544 -0.612 -0.925

Standardized residuals:
     Min       Q1      Med       Q3      Max
-2.33375 -0.61913  0.09387  0.69813  2.69217

Residual standard error: 2.579
Degrees of freedom: 252 total; 248 residual
```

d)

Consideramos o ajuste `ajuste_gee_unstructured` (Não Estruturada), como o mais adequado aos dados. abaixos temos o resultado do `summary`do modelo.

```r
summary(ajuste_gee_unstructured)$coefficients
```

```
            Estimate Std.err    Wald Pr(>|W|)
(Intercept) 70.55828 2.35043 901.157  0.00000
group       -2.50386 1.05934   5.587  0.01810
time         0.08166 0.04398   3.448  0.06335
group:time   0.03421 0.01848   3.428  0.06410
```

Com base nos resultados, podemos observar que as variáveis `time` e interação de tempo e grupo (`group * time`), não são significativas a um nível de 5%. Mas podemos ver que o efeito de grupo é significativo.

Questão 3

a)

```r
ajuste_misto_intercept <- lme(response ~ group*time, random = ~ 1 | subject, data = dados_ra

ajuste_misto_tempo <- lme(response ~ group * time,  random = ~ 1 + time | subject, data = da
```

Ambos os modelos ajustado assumem que \$ Y_{ij}|b_i N(,)\$ e que os $b_i(b_{0i}, b_{1i}) \sim N(0, \Sigma)$. Onde b_{0i} é o efeito aleatório do intercepto e b_{1i} do tempo.

b)

Para o ajuste com apenas o intercepto aleatório (ajuste_misto_intercept), interceptos diferentes geram correlações constantes entre todas observações do mesmo indivíduo.

Já para o modelo com efeito aleatório no intercepto e no tempo, a correlação intraindivíduo varia ao longo do tempo.

c)

Modelo com intercepto aleatório: São estimados 3 parâmetros (variância do intercepto, variância residual e a média fixada).

Modelo com intercepto e tempo aleatórios: São estimados 5 parâmetros (variâncias do intercepto, tempo, covariância entre eles e variância residual).

Modelos marginais (GEE): O número de parâmetros depende da estrutura de correlação. As estruturas mais complexas, como a não estruturada, estimam mais parâmetros.

Ao dobrar as medidas repetidas, os modelos mistos mantém o mesmo número de parâmetros. Já o GEE, como o não estruturado, por exemplo, aumenta quadraticamente o número de parâmetros a serem estimados. Sendo assim, os modelos mistos acomodam melhor um número maior de medidas repetidas.

d)

Com base no menor AIC e BIC, o modelo que parace se ajustar melhor aos dados é o com efeito aleatório apenas no intercepto.

modelo	AIC	BIC
ajuste_misto_intercept	1098	1119
ajuste_misto_tempo	1102	1130

e)

Podemos apresentar os efeitos aleatórios por meio da função `ranef()`, abaixo temos os resultados:

```
head(ranef(ajuste_misto_intercept))
```

```
  (Intercept)
```

```
1     -2.4239
3      2.6289
5     -2.6379
6      3.2329
7      0.0938
8      0.9159
```

As estimativas da variância condicional podem ser encontradas como:

```r
getVarCov(ajuste_misto_intercept, type = 'conditional')
```

```
subject 8
Conditional variance covariance matrix
      1     2
1 2.792 0.000
2 0.000 2.792
  Standard Deviations: 1.671 1.671
```

f)

Com base no `summary`do modelo selecionado, podemos observar com base no p-valor associado, que não parece ter um efeito significativo de grupo, nem na relação de tempo e grupo.

```r
summary(ajuste_misto_intercept)
```

```
Linear mixed-effects model fit by REML
  Data: dados_rats
   AIC  BIC logLik
  1098 1119 -542.8

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:       1.951    1.671

Fixed effects:  response ~ group * time
            Value Std.Error  DF t-value p-value
(Intercept) 63.80    1.3541 200   47.11  0.0000
group       -0.34    0.6632  48   -0.52  0.6062
time         0.19    0.0156 200   12.23  0.0000
```

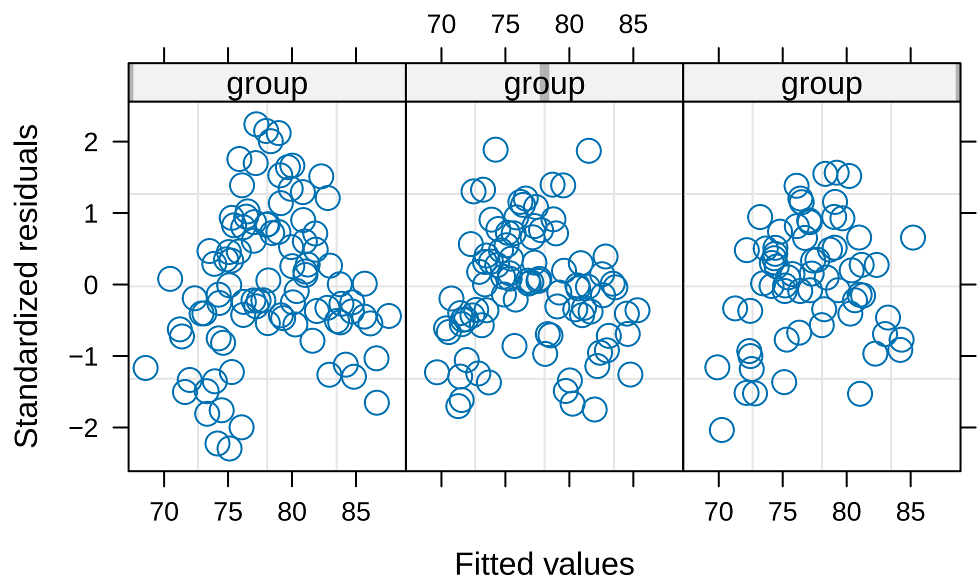
```
group:time   0.00    0.0078 200    0.22  0.8284
 Correlation:
           (Intr) group  time
group      -0.923
time       -0.818  0.771
group:time  0.749 -0.827 -0.924

Standardized Within-Group Residuals:
      Min         Q1        Med         Q3        Max
-2.294210 -0.546587 -0.003812  0.653397  2.242103

Number of Observations: 252
Number of Groups: 50
```

g)

Abaixo temos um gráfico de resíduos padronizados versus os valores preditos para o grupo.



Os redíduos se distribuem ao redor de zero com uma amplitude pequena no eixo y dos gráficos, indicando um bom ajuste do modelo aos dados nos diferentes grupos.

h)

A escolha do modelo depende do contexto da análise. O modelo misto é ideal quando o objetivo é capturar a variação intraindivíduo, enquanto o modelo marginal (GEE) foca nas diferenças entre grupos ou tratamentos aplicados.

De modo geral, o modelo marginal (GEE) é preferido, pois os coeficientes são mais fáceis de interpretar e ele é aplicável em uma ampla gama de situações. Já o modelo misto deve ser utilizado em estudos onde é essencial considerar a variação intraindivíduo como parte do objetivo principal da análise.