

INSTITUTO SUPERIOR DE AGRONOMIA
Modelos Matemáticos e Aplicações (2020-21)

Teste – Modelos Lineares Generalizados e Modelos Lineares Mistos

31 de Maio, 2021

Duração: 2h30

I [9 valores]

Um estudo visa estimar o número de bagos por cacho de uvas (variável de contagem BE, do inglês *berries*) a partir de três outras variáveis: o peso do cacho (variável Bw, de *bunch weight*, em *g*) e duas variáveis observáveis em imagens a duas dimensões obtidas por robôs que percorrem a vinha, a saber, o número de bagos por cacho visíveis nas imagens (variável BEv, de contagem) e a área do cacho nas imagens (variável Ba, de *bunch area*, em cm^2). O conjunto de dados usado para ajustar o modelo incluía observações sobre 75 cachos de cada uma de 5 castas, num total de 375 observações, mas como se pretendia um modelo a aplicar a qualquer casta, as observações foram tratadas em conjunto.

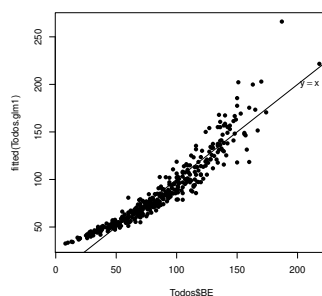
Eis alguns indicadores de síntese:

```
> summary(Todos[,c("BE", "BEv", "Ba", "Bw")])
      BE      BEv      Ba      Bw
Min.   : 8.0   Min.   : 8.0   Min.   : 10.60  Min.   : 10.6
1st Qu.: 61.0  1st Qu.: 34.0  1st Qu.: 54.52  1st Qu.: 86.0
Median : 85.0  Median : 44.0  Median : 74.12  Median : 133.6
Mean   : 87.7  Mean   : 44.7  Mean   : 74.44  Mean   : 137.2
3rd Qu.: 113.5 3rd Qu.: 55.0  3rd Qu.: 90.67  3rd Qu.: 174.8
Max.   : 218.0 Max.   : 83.0  Max.   : 154.62  Max.   : 351.0
```

1. Tendo em conta a natureza da componente aleatória BE diga, justificando, qual a distribuição de probabilidades mais adequada para a descrever (de entre as estudadas nas aulas).
2. Independentemente da resposta na alínea anterior, foram ajustados dois Modelos Lineares Generalizados com variável resposta Poisson, mas diferindo na sua função de ligação. Eis os resultados:

<pre>> summary(Todos.glm1) Call: glm(formula = BE ~ BEv + Bw + Ba, family = poisson(link = log), data = Todos) Coefficients: Estimate Std. Error z value Pr(> z) (Intercept) 3.3338620 0.0201937 165.094 <2e-16 BEv 0.0166706 0.0007675 21.721 <2e-16 Bw 0.0029365 0.0001973 14.881 <2e-16 Ba -0.0011815 0.0005132 -2.302 0.0213 --- Null deviance: 5970.14 on 374 degrees of freedom Residual deviance: 676.76 on 371 degrees of freedom AIC: 3012.5</pre>	<pre>> summary(Todos.glm2) Call: glm(formula = BE ~ BEv + Bw + Ba, family = poisson(link = identity), data = Todos) Coefficients: Estimate Std. Error z value Pr(> z) (Intercept) -4.07425 1.14234 -3.567 0.000362 BEv 1.38669 0.07167 19.348 < 2e-16 Bw 0.34403 0.02071 16.613 < 2e-16 Ba -0.23393 0.05079 -4.606 4.1e-06 --- Null deviance: 5970.14 on 374 degrees of freedom Residual deviance: 267.95 on 371 degrees of freedom AIC: 2603.7</pre>
--	--

- (a) Descreva em pormenor o modelo indicado à esquerda (modelo `Todos.glm1`).
- (b) Eis a nuvem de pontos correspondentes aos bagos por cacho (eixo horizontal) e correspondentes valores estimados pelo modelo `Todos.glm1` (à esquerda), com a recta $y = x$. Comente-o.



- (c) Indique o número médio de bagos que o modelo ajustado à direita (modelo `Todos.glm2`) faria corresponder a um cacho de peso 20 g e com, na respectiva imagem, 15 cm² de área e 10 bagos visíveis. Comente, sabendo também que o valor previsto pelo outro modelo é 34.521.
- (d) Qual dos dois modelos escolheria, com base na informação disponível? Justifique a sua resposta.
- (e) Considere uma modificação ao modelo `Todos.glm2` (à direita): considerar que a distribuição da sua componente aleatória é Normal. Comente esse modelo. Como seria possível comparar este novo modelo com o modelo `Todos.glm2`?
3. Os modelos acima ajustados incluem um preditor que exige a pesagem manual dos bagos (`Bw`). Procurou-se um modelo cuja componente sistemática dependa apenas de medições sobre as imagens recolhidas de forma automática. Foi ajustado um modelo Poisson, com função ligação identidade, e apenas dois preditores: `BEv` e `Ba`. O desvio residual obtido foi 547.3. Efectue um Teste de Razão de Verosimilhanças para saber se este novo modelo tem uma qualidade de ajustamento significativamente pior do que o modelo correspondente com 3 preditores. Comente.

II [11 valores]

1. Com vista a estudar a variabilidade genética do rendimento (kg/planta) em clones da variedade de oliveira Cobrançosa nos primeiros anos de intalação do olival, avaliaram-se 125 clones quanto a essa característica num ensaio com um delineamento experimental em blocos casualizados completos (5 blocos). Em cada bloco existe apenas uma observação por clone. Admita que, tanto o `bloco` como o `clone`, são factores de efeitos aleatórios.
- (a) Descreva, em pormenor, o modelo que lhe parece ser adequado para o estudo acima descrito.
- (b) Em notação matricial, descreva os pressupostos do modelo que definiu na alínea a).
- (c) No R, com a função `lmer` do pacote `lme4` executaram-se os seguintes comandos:

```
> library(lme4)
> library(lmerTest)
> dadoslmer1<-lmer(rend~1+(1|clone)+(1|bloco), data=dados)
> summary(dadoslmer1)
Linear mixed model fit by REML.
t-tests use Satterthwaite's method [lmerModLmerTest]
Formula: rend ~ 1 + (1 | clone) + (1 | bloco)
Data: dados
REML criterion at convergence: 698.7
Random effects:
  Groups   Name                Variance Std.Dev.
clone     (Intercept)  0.04020  0.2005
bloco     (Intercept)  0.01124  0.1060
Residual                    0.14741  0.3839
Number of obs: 625, groups:  clone, 125;  bloco, 5
Fixed effects:
              Estimate Std. Error    df t value Pr(>|t|)
(Intercept)  0.55415     0.05296  5.09465   10.46 0.000123 ***
---
> logLik(dadoslmer1)
'log Lik.' -349.342
> dadoslmer2<-lmer(rend~1+(1|clone), data=dados)
> logLik(dadoslmer2)
'log Lik.' -362.997
> dadoslmer3<-lmer(rend~1+(1|bloco), data=dados)
> logLik(dadoslmer3)
'log Lik.' -370.7699
> ranef(dadoslmer1)
$clone
```

```

      (Intercept)
CB1011 -0.083164783
CB1013  0.004873856
CB1021 -0.067703213
CB1023 -0.145472601
CB1024  0.239105097
CB1031  0.013412335
...

$bloco
  (Intercept)
B1 -0.15542972
B2  0.06774529
B3  0.10288132
B4  0.01713723
B5 -0.03233413

```

- i. Efectue os testes às componentes de variância que achar convenientes de acordo com o modelo descrito anteriormente. Descreva detalhadamente apenas um dos testes de hipóteses que efectuar.
 - ii. De acordo com o Critério de Informação de Akaike (AIC) qual dos 3 modelos ajustados é o melhor.
- (d) De acordo com o modelo completo ajustado, qual o rendimento previsto para o genótipo CB1011 no bloco B1?
2. Um investigador defende que, dado o reduzido número de níveis do factor `bloco`, seria defensável admiti-lo como um factor de efeitos fixos. O ajustamento deste modelo no R, com a função `lmer` do pacote `lme4` proporcionou os seguintes resultados:

```

> dadoslmer4<-lmer(rend~bloco+(1|clone), data=dados)
> summary(dadoslmer4)
Linear mixed model fit by REML.
t-tests use Satterthwaite's method [lmerModLmerTest]
Formula: rend ~ bloco + (1 | clone)
Data: dados
REML criterion at convergence: 703.3
Random effects:
Groups   Name          Variance Std.Dev.
clone    (Intercept)  0.0402  0.2005
Residual                    0.1474  0.3839
Number of obs: 625, groups: clone, 125
Fixed effects:
      Estimate Std. Error    df t value Pr(>|t|)
(Intercept)  0.38241    0.03874 523.79227  9.871 < 2e-16 ***
blocoB2      0.24660    0.04857 496.00000  5.078 5.41e-07 ***
blocoB3      0.28542    0.04857 496.00000  5.877 7.67e-09 ***
blocoB4      0.19068    0.04857 496.00000  3.926 9.85e-05 ***
blocoB5      0.13602    0.04857 496.00000  2.801  0.0053 **

```

- (a) Defina a covariância entre observações feitas no mesmo bloco para o modelo que admite o `bloco` como um factor de efeitos fixos e para o modelo que admite o `bloco` como um factor de efeitos aleatórios. Interprete os resultados obtidos.
- (b) Sabendo que $\bar{y}_{..} = 0.554$ kg/planta e que $\bar{y}_{CB12} = 0.969$ kg/planta, qual o melhor preditor empírico linear não enviesado (EBLUP) do rendimento do clone CB12? Explique o seu significado.
- (c) O rendimento obtido no bloco 2 é significativamente diferente do rendimento obtido no bloco 1? Justifique a sua resposta.