

---

INSTITUTO SUPERIOR DE AGRONOMIA  
**Modelos Matemáticos e Aplicações– 2019-20**  
**Algumas resoluções de Exercícios de Modelos Lineares Generalizados**

1. Ver acetatos das aulas
2. FALTA
3. Para aceder aos dados é necessário primeiro carregar o módulo (*package*) MASS. Eis esse carregamento e as primeiras linhas da *data frame* Traffic:

```
> library(MASS)
> head(Traffic)
year day limit  y
1 1961  1    no  9
2 1961  2    no 11
3 1961  3    no  9
4 1961  4    no 20
5 1961  5    no 31
6 1961  6    no 26
```

- (a) A componente aleatória deste modelo é a variável  $Y$  que conta o número de acidentes por dia. É uma variável de contagem à qual é natural associar uma distribuição de Poisson, com parâmetro desconhecido  $\lambda$  (que é também o valor esperado de  $Y$ ). Admite-se que o valor de  $\lambda$  depende da única variável preditora (**limit**), uma variável categórica muito simples: tem apenas dois níveis, definidos pela existência, ou não, do limite de velocidade máxima nas estradas suecas. Finalmente, e tratando-se dum modelo loglinear, a função de ligação é a função logarítmica, ou seja,  $g(\lambda) = \ln(\lambda)$ . Dada a natureza do problema, utilizaremos uma dupla indexação para designar as observações da componente aleatória  $Y$ , sendo um primeiro índice  $i = 1, 2$  associado aos dois níveis do factor (que pela convenção do R ficam ordenados pela ordem alfabética dos seus nomes em inglês:  $i=1$  corresponde a “no” e  $i=2$  a “yes”), e um segundo índice  $j$  indica as repetições em cada nível do factor. Como se pode verificar inspeccionando a *data frame*,  $j$  varia de 1 a 115 se  $i=1$  e de 1 a 69 se  $i=2$ :

```
> summary(Traffic)
year      day      limit      y
Min.   :1961  Min.   : 1.00  no :115  Min.   : 7.00
1st Qu.:1961  1st Qu.:23.75  yes: 69  1st Qu.:15.00
Median :1962  Median :46.50             Median :20.00
Mean   :1962  Mean   :46.50             Mean   :21.55
3rd Qu.:1962  3rd Qu.:69.25             3rd Qu.:26.00
Max.   :1962  Max.   :92.00             Max.   :49.00
```

Assim, a equação do modelo é:

$$\begin{aligned}\ln(\lambda_{ij}) &= g(E[Y_{ij}]) = \mu + \alpha_i \\ \Leftrightarrow \lambda_{ij} &= e^\mu \cdot e^{\alpha_i},\end{aligned}$$

sendo  $\alpha_1=0$  e  $\alpha_2$  o parâmetro associado ao efeito do limite de velocidade.

Por outras palavras, pelo modelo o valor do parâmetro  $\lambda$  apenas pode depender do nível do factor, e o índice  $i$  identifica os dois possíveis valores de  $\lambda$ :  $\lambda_1 = e^\mu$  é o número médio de acidentes por dia quando não existia limite de velocidade, e  $\lambda_2 = e^\mu e^{\alpha_2}$  é o número médio

de acidentes por dia após a introdução do limite de velocidade. Identificando directamente o significado de cada parâmetro, tem-se  $\mu = \ln(\lambda_1)$ , ou seja, o logaritmo do valor esperado de acidentes por dia quando não há limite de velocidade, e  $\alpha_2 = \ln\left(\frac{\lambda_2}{\lambda_1}\right) = \ln(\lambda_2) - \ln(\lambda_1)$  é a diferença dos logaritmos desses valores médios, nos dias com e sem, respectivamente, limite de velocidade. É de esperar que  $\alpha_2$  seja negativo.

Ajustemos o modelo, através dos seguintes comandos R:

```
> Traff.glm <- glm(y ~ limit, family=poisson, data=Traffic)
> summary(Traff.glm)
[...]
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.14115    0.01939 162.006 < 2e-16 ***
limityes     -0.20130    0.03380  -5.956 2.58e-09 ***
---
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 625.25  on 183  degrees of freedom
Residual deviance: 589.00  on 182  degrees of freedom
AIC: 1483
Number of Fisher Scoring iterations: 4
```

Os valores estimados dos parâmetros do modelo são os seguintes:  $\hat{\mu} = 3.14115$  e  $\hat{\alpha}_2 = -0.20130$ . Assim, o número esperado de acidentes por dia estimado pelo modelo é, em dias sem limite de velocidade,  $\hat{\lambda}_1 = e^{3.14115} = 23.13045$ . A média estimada correspondente, para dias com limite de velocidade, é  $\hat{\lambda}_2 = \hat{\lambda}_1 e^{\hat{\alpha}_2} = 23.13045 \times e^{-0.20130} = 18.91301$ . Assim, a média de acidentes por dia estimada a partir do modelo sofre uma redução multiplicativa de  $e^{-0.20130} = 0.8176671$ , com a introdução do limite de velocidade. Pode dizer-se que, em média, o número esperado de acidentes por dia cai cerca de 18%, quando está em vigor o limite de velocidade.

- (b) Nesta alínea pede-se para calcular o número médio de acidentes por dia, para cada nível do factor, não a partir do modelo, mas apenas como média das observações existentes. Eis uma possível maneira de obter esses valores no R, usando o comando `by`:

```
> by(Traffic$y, Traffic$limit, FUN=mean)
Traffic$limit: no
[1] 23.13043
-----
Traffic$limit: yes
[1] 18.91304
```

O facto mais saliente dos valores obtidos é que (erros de arredondamento aparte) as médias amostrais coincidem com os valores médios estimados a partir do modelo log-linear. Surge naturalmente a pergunta de se estamos perante uma coincidência, ou um resultado obrigatório, questão que é abordada na alínea seguinte.

- (c) **[Material Complementar:]** A função verosimilhança de  $n$  observações, no contexto refe-

rido, é dada por:

$$\begin{aligned}
 L(\lambda_i) &= \prod_{i,j} f(\lambda_i; y_{ij}) = \prod_{i,j} \frac{e^{-\lambda_i} \lambda_i^{y_{ij}}}{y_{ij}!} = \prod_{i,j} e^{-\lambda_i + y_{ij} \ln(\lambda_i) - \ln(y_{ij}!)} \\
 &= e^{\sum_{i,j} [-\lambda_i + y_{ij} \ln(\lambda_i) - \ln(y_{ij}!)]} \\
 \Leftrightarrow \mathcal{L}(\lambda_i) &= \ln(L(\lambda_i)) = \sum_{i,j} [-\lambda_i + y_{ij} \ln(\lambda_i) - \ln(y_{ij}!)]
 \end{aligned}$$

Re-escrevendo esta log-verosimilhança como função dos dois parâmetros do modelo ( $\mu$  e  $\alpha_2$ ), e utilizando a já referida relação  $\lambda_i = e^{\mu + \alpha_i}$  (com  $\alpha_1 = 0$ ), tem-se:

$$\mathcal{L}(\lambda_i) = \sum_{i,j} [-e^{\mu + \alpha_i} + y_{ij}(\mu + \alpha_i) - \ln(y_{ij}!)] .$$

Separando o somatório em dois somatórios, associados respectivamente a  $i = 1$  (com  $n_1$  parcelas associadas aos dias sem limite de velocidade) e  $i = 2$  (com  $n_2$  parcelas associadas aos dias com limite de velocidade), vem:

$$\mathcal{L}(\mu, \alpha_i) = \sum_{j=1}^{n_1} [-e^{\mu} + y_{1j}\mu - \ln(y_{1j}!)] + \sum_{j=1}^{n_2} [-e^{\mu + \alpha_2} + y_{2j}(\mu + \alpha_2) - \ln(y_{2j}!)]$$

Logo, sendo  $\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^2 \sum_{j=1}^{n_i} y_{ij}$  e  $\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$  as médias global, e das observações no nível  $i$  do factor, respectivamente, tem-se que as derivadas parciais da log-verosimilhança em relação a cada parâmetro são dadas por:

$$\begin{aligned}
 \frac{\partial \mathcal{L}}{\partial \mu} &= \sum_{j=1}^{n_1} [-e^{\mu} + y_{1j}] + \sum_{j=1}^{n_2} [-e^{\mu + \alpha_2} + y_{2j}] = -n_1 e^{\mu} - n_2 e^{\mu + \alpha_2} + n \bar{y}_{..} \\
 \frac{\partial \mathcal{L}}{\partial \alpha_2} &= \sum_{j=1}^{n_2} [-e^{\mu + \alpha_2} + y_{2j}] = -n_2 e^{\mu + \alpha_2} + n_2 \bar{y}_{2.}
 \end{aligned}$$

Anulando o vector gradiente, obtem-se o seguinte sistema que, neste caso admite solução analítica:

$$\begin{aligned}
 \left\{ \begin{array}{l} \frac{\partial \mathcal{L}}{\partial \mu} = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_2} = 0 \end{array} \right. &\Leftrightarrow \left\{ \begin{array}{l} \bar{y}_{..} = \frac{n_1 e^{\hat{\mu} + n_2 e^{\hat{\mu} + \hat{\alpha}_2}}}{n} \\ \bar{y}_{2.} = e^{\hat{\mu} + \hat{\alpha}_2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \bar{y}_{..} = \frac{n_1 e^{\hat{\mu} + n_2 \bar{y}_{2.}}}{n} \\ \ln(\bar{y}_{2.}) = \hat{\mu} + \hat{\alpha}_2 \end{array} \right. \\
 &\Leftrightarrow \left\{ \begin{array}{l} \hat{\mu} = \ln\left(\frac{n \bar{y}_{..} - n_2 \bar{y}_{2.}}{n_1}\right) \\ \hat{\mu} + \hat{\alpha}_2 = \ln(\bar{y}_{2.}) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \hat{\mu} = \ln(\bar{y}_{1.}) \\ \hat{\alpha}_2 = \ln(\bar{y}_{2.}) - \ln(\bar{y}_{1.}) \end{array} \right. ,
 \end{aligned}$$

já que  $\bar{y}_{1.} = \frac{n \bar{y}_{..} - n_2 \bar{y}_{2.}}{n_1}$  (ver as definições das médias em causa). Assim, as estimativas de máxima verosimilhança dos parâmetros  $\mu = \ln(\lambda_1)$  e  $\alpha_2 = \ln(\lambda_2) - \ln(\lambda_1)$  são obtidas substituindo os valores esperados  $\lambda_1$  e  $\lambda_2$  pelas respectivas médias amostrais  $\bar{y}_{1.}$  e  $\bar{y}_{2.}$ . O resultado numérico da alínea anterior não é uma coincidência.

- (d) Tendo em conta as alíneas anteriores, pode afirmar-se que a diferença principal no estudo da hipótese de que  $\lambda_1 = \lambda_2$  (que no caso do modelo log-linear corresponde a um teste a

$\alpha_2 = 0$ ), nas duas abordagens referidas, está associada às hipóteses distribucionais sobre a componente aleatória  $Y$ . Nos testes clássicos de comparação de médias admite-se a Normalidade (e homogeneidade de variâncias) de  $Y$ . Se  $Y$  é uma variável de contagem (como neste exemplo), essa hipótese de Normalidade apenas pode ser aproximadamente verdadeira. Já a hipótese de uma distribuição de Poisson, usada no modelo log-linear deste exercício, é mais adequada à natureza da variável de contagem  $Y$ .

4. O conjunto de dados referido no enunciado é o seguinte. As linhas correspondem a locais e as colunas a espécies.

```
> waders
```

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19
A	12	2027	0	0	2070	39	219	153	0	15	51	8336	2031	14941	19	3566	0	5	0
B	99	2112	9	87	3481	470	2063	28	17	145	31	1515	1917	17321	3378	20164	177	1759	53
C	197	160	0	4	126	17	1	32	0	2	9	477	1	548	13	273	0	0	0
D	0	17	0	3	50	6	4	7	0	1	2	16	0	0	3	69	1	0	0
E	77	1948	0	19	310	1	1	64	0	22	81	2792	221	7422	10	4519	12	0	0
F	19	203	48	45	20	433	0	0	11	167	12	1	0	26	1790	2916	473	658	55
G	1023	2655	0	18	320	49	8	121	9	82	48	3411	14	9101	43	3230	587	10	5
H	87	745	1447	125	4330	789	228	529	289	904	34	1710	7869	2247	4558	40880	7166	1632	498
I	788	2174	0	19	224	178	1	423	0	195	162	2161	25	1784	3	1254	0	0	0
J	82	350	760	197	858	962	10	511	251	987	191	34	87	417	4496	15835	5327	1312	1020
K	474	930	0	10	316	161	0	90	0	39	48	1183	166	4626	65	127	4	0	0
L	77	249	160	136	999	645	15	851	101	723	266	495	83	1253	1864	4107	1939	623	527
M	22	144	0	4	1	1	0	10	0	2	9	125	5	411	0	3	0	0	0
N	0	791	0	0	4	38	1	56	1	30	54	95	0	1726	0	0	0	0	0
O	0	360	128	43	364	1628	63	287	328	641	850	83	67	48	6499	9094	5647	1333	582

- (a) O teste  $\chi^2$  pedido (estudado no módulo I desta UC) tem por hipótese nula a independência dos dois factores de classificação (local e espécie). Ou seja, tem por hipótese nula que, para qualquer célula (combinação local/espécie)  $(i, j)$ , se verifica que a probabilidade conjunta de estar nessa célula é o produto das probabilidades marginais de estar nesse local e nessa espécie:  $\pi_{ij} = \pi_{i.} \times \pi_{.j}$ ,  $\forall i, j$ . Sabe-se que a estatística desse teste (estatística de Pearson),

é  $X^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$ , onde  $O_{ij}$  indica o número de indivíduos observados no local  $i =$

$1, \dots, a$ , da espécie  $j = 1, \dots, b$ , e  $\hat{E}_{ij}$  indica o correspondente valor esperado, estimado ao abrigo da hipótese nula de independência. Essa estimativa é dada por  $\hat{E}_{ij} = N \times \frac{N_{i.}}{N} \times \frac{N_{.j}}{N} = \frac{N_{i.} \times N_{.j}}{N}$ , sendo  $N$  o número total de observações,  $N_{i.}$  o número total de observações no local  $i$  e  $N_{.j}$  o número total de observações da espécie  $j$ . Caso seja verdade a hipótese nula de independência, esta estatística tem, assintoticamente, distribuição  $\chi^2_{(a-1)(b-1)}$ . Neste caso, o resultado do teste obtém-se pelo comando do R indicado no enunciado:

```
> chisq.test(waders)
Pearson's Chi-squared test
data: waders
X-squared = 248356, df = 252, p-value < 2.2e-16
Warning message:
In chisq.test(waders) : Chi-squared approximation may be incorrect
```

O  $p$ -value próximo de zero associado ao valor calculado da estatística indica uma claríssima rejeição da hipótese nula de independência (ao abrigo da qual haviam sido obtidos os valores esperados estimados  $\hat{E}_{ij}$ ). Este resultado era de esperar, dada a natureza dos dados. De facto, é de esperar que haja relações de preferência ou aversão de determinadas espécies

por determinados locais, ao contrário do que seria o caso sob a hipótese de independência. **Nota:** A advertência final resulta do facto de a distribuição ser apenas assintótica, sendo habitual considerar que a aproximação é válida caso não existam células com um número muito reduzido de observações esperadas (os critérios exactos diferem). Neste caso, existem várias células com valores pequenos de  $\hat{E}_{ij}$ , o que motiva a advertência do R.

- (b) Eis a cabeça da *data frame* criada com o comando do enunciado:

```
> head(limicolas)
obs local especie
1  12     A      S1
2  99     B      S1
3 197     C      S1
4   0     D      S1
5  77     E      S1
6  19     F      S1
```

- (c) Neste contexto, a componente aleatória do MLG pedido é a variável  $Y$  que conta o número de observações em cada célula. Embora existam dois factores (**local**, com  $a = 15$  níveis e **especie**, com  $b = 19$  níveis), existe uma única contagem por célula, pelo que apenas são necessários dois índices para identificar cada observação. Dada a natureza da variável resposta, é natural admitir que cada uma das  $n = ab = 15 \times 19 = 285$  observações  $Y_{ij}$  corresponda a uma concretização duma distribuição de Poisson, sendo o parâmetro da distribuição específico para cada célula, ou seja, o parâmetro da Poisson pode variar consoante o local e a espécie:  $\lambda_{ij}$ . Desta forma, o número esperado de observações em cada célula é dado por  $\lambda_{ij} = E[Y_{ij}]$ . Uma vez que não existem repetições nas  $ab$  células deste delineamento factorial a dois factores, não é possível prever efeitos de interacção (fazê-lo corresponderia a trabalhar com um modelo saturado, com tantas observações quantos os parâmetros do modelo:  $ab$ ). Assim, a parte sistemática corresponderá à de um modelo a dois factores sem efeitos de interacção:  $\mu + \alpha_i + \beta_j$ , sendo  $\alpha_i$  o efeito do local  $i$  ( $\alpha_1 = 0$ ) e  $\beta_j$  o efeito da espécie  $j$  ( $\beta_1 = 0$ ). Finalmente, a função de ligação deste MLG é o logaritmo. Assim, a equação de base deste modelo log-linear é da forma:

$$g(E[Y_{ij}]) = \ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j .$$

O valor esperado na célula  $(i, j)$  é dado por

$$\lambda_{ij} = e^{\mu + \alpha_i + \beta_j} ,$$

(com  $\alpha_1 = \beta_1 = 0$ ). No caso concreto da espécie  $S14$  ( $j = 14$ ) e local C ( $i = 3$ ), esse valor esperado é dado por

$$\lambda_{3,14} = e^{\mu + \alpha_3 + \beta_{14}} .$$

- (d) O modelo pedido no enunciado é o seguinte:

```
> summary(limic.glm)
Call: glm(formula = obs ~ local + especie, family = poisson, data = limicolas)
[...]
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.767930   0.019100  301.991 < 2e-16 ***
localB       0.493097   0.006936   71.095 < 2e-16 ***
localC      -2.890491   0.023822 -121.336 < 2e-16 ***
```

---

```

localD      -5.231437   0.074943  -69.806 < 2e-16 ***
localE      -0.648924   0.009328  -69.568 < 2e-16 ***
localF      -1.582885   0.013239 -119.560 < 2e-16 ***
localG      -0.479293   0.008837  -54.236 < 2e-16 ***
localH       0.820547   0.006558  125.116 < 2e-16 ***
localI      -1.271316   0.011677 -108.874 < 2e-16 ***
localJ       0.006044   0.007717   0.783  0.43348
localK      -1.402189   0.012298 -114.018 < 2e-16 ***
localL      -0.795512   0.009800  -81.178 < 2e-16 ***
localM      -3.816235   0.037239 -102.481 < 2e-16 ***
localN      -2.482878   0.019685 -126.128 < 2e-16 ***
localO      -0.177257   0.008095  -21.898 < 2e-16 ***
especieS10   0.290805   0.024311  11.962 < 2e-16 ***
especieS11  -0.470071   0.029653  -15.853 < 2e-16 ***
especieS12   2.026402   0.019564  103.580 < 2e-16 ***
especieS13   1.440433   0.020451   70.433 < 2e-16 ***
especieS14   3.040876   0.018824  161.547 < 2e-16 ***
especieS15   2.039994   0.019548  104.357 < 2e-16 ***
especieS16   3.579613   0.018644  192.000 < 2e-16 ***
especieS17   1.976080   0.019622  100.705 < 2e-16 ***
especieS18   0.908073   0.021784   41.685 < 2e-16 ***
especieS19  -0.076217   0.026517   -2.874  0.00405 **
especieS2    1.614834   0.020135   80.199 < 2e-16 ***
especieS3   -0.147298   0.027019   -5.452  4.99e-08 ***
especieS4   -1.426666   0.041793  -34.137 < 2e-16 ***
especieS5    1.516512   0.020307   74.678 < 2e-16 ***
especieS6    0.605367   0.022864   26.477 < 2e-16 ***
especieS7   -0.123294   0.026846   -4.593  4.38e-06 ***
especieS8    0.067029   0.025582    2.620  0.00879 **
especieS9   -1.077200   0.036486  -29.524 < 2e-16 ***

```

---

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1039396 on 284 degrees of freedom

Residual deviance: 251164 on 252 degrees of freedom

AIC: 252821

Number of Fisher Scoring iterations: 6

Uma primeira análise do desvio residual diz-nos que é apenas cerca de 1/4 do desvio do modelo nulo, pelo que a utilização de local e espécie como factores preditores explica parte importante das contagens observadas.

A contagem esperada ao abrigo do modelo na célula do local C, espécie S14, é

$$\hat{\lambda}_{3,14} = e^{\hat{\mu} + \hat{\alpha}_3 + \hat{\beta}_{14}} = e^{5.767930 - 2.890491 + 3.040876} = 371.7847.$$

A contagem efectiva nessa célula foi  $O_{3,14} = 548$ . A discrepância é substancial, embora a ordem de grandeza da contagem esteja certa.

- (e) A soma de quadrados dos resíduos de Pearson é, neste contexto, igual ao valor da estatística do teste  $\chi^2$  de Pearson. Este facto torna-se evidente, tendo em conta as definições dos resíduos de Pearson (ver acetatos MLGs) e da estatística do teste  $\chi^2$  (ver acima).

```

> sum(residuals(limic.glm, type="pearson")^2)
[1] 248356

```

Assim, pode pensar-se no valor da estatística de Pearson no teste  $\chi^2$  à independência, como uma medida do afastamento dos valores esperados estimados pelo modelo log-linear considerado (sem efeitos de interação), em relação aos valores efectivamente observados, ou seja, como uma medida do mau ajustamento deste modelo.

- (f) Para uma mesma espécie, o quociente dos valores esperados de  $Y$  nos locais 4 e 3 é:

$$\frac{\lambda_{4j}}{\lambda_{3j}} = \frac{e^{\mu+\alpha_4+\beta_j}}{e^{\mu+\alpha_3+\beta_j}} = e^{\alpha_4-\alpha_3}.$$

Logo,  $\alpha_4 - \alpha_3 = \ln\left(\frac{\lambda_{4j}}{\lambda_{3j}}\right) = \ln(\lambda_{4j}) - \ln(\lambda_{3j})$ , pelo que se trata da diferença entre os logaritmos dos valores esperados das contagens, duma mesma espécie, nos locais C e D.

- (g) Conhecemos os intervalos de confiança assintóticos duma qualquer combinação linear dos parâmetros. Para esta combinação linear simples, temos a seguinte expressão dum IC a 95% de confiança (ver acetatos de MLGs):

$$\left] (\hat{\alpha}_4 - \hat{\alpha}_3) - z_{0.025} \cdot \hat{\sigma}_{\hat{\alpha}_4 - \hat{\alpha}_3}, (\hat{\alpha}_4 - \hat{\alpha}_3) + z_{0.025} \cdot \hat{\sigma}_{\hat{\alpha}_4 - \hat{\alpha}_3} \right[.$$

Os valores estimados de  $\alpha_4$  e  $\alpha_3$  são dados na listagem:  $\hat{\alpha}_4 = -5.231437$  e  $\hat{\alpha}_3 = -2.890491$ , pelo que  $\hat{\alpha}_4 - \hat{\alpha}_3 = -2.340946$ . Para calcular o erro padrão estimado associado à diferença  $\hat{\alpha}_4 - \hat{\alpha}_3$ , necessitamos da matriz de variâncias-covariâncias estimadas dos parâmetros, ou seja da inversa da matriz de informação de Fisher (ver acetato dos MLGs). Dada a sua dimensão neste exemplo, optamos por apenas apresentar as linhas e colunas desta matriz correspondentes aos parâmetros  $\alpha_4$  e  $\alpha_3$ :

```
> vcov(limic.glm)[c("localC","localD"),c("localC","localD")]
localC      localD
localC 5.674991e-04 2.986499e-05
localD 2.986499e-05 5.616452e-03
```

Assim, tem-se:

$$\begin{aligned}\hat{\sigma}_{\hat{\alpha}_4 - \hat{\alpha}_3} &= \sqrt{\hat{V}[\hat{\alpha}_4] + \hat{V}[\hat{\alpha}_3] - 2 \cdot Cov[\hat{\alpha}_4, \hat{\alpha}_3]} \\ &= \sqrt{0.005616452 + 0.0005674991 - 2(0.00002986499)} \\ &= \sqrt{0.006124221} = 0.0782574.\end{aligned}$$

Como  $z_{0.025} = 1.959964$ , o IC pedido é:

$$\left] -3.043873, -2.737109 \right[.$$

A diferença entre os logaritmos dos valores esperados das contagens, duma mesma espécie, nos locais C e D, está contida neste IC, com 95% de confiança. Ou seja, a razão desses valores esperados está, a 95% de confiança, entre  $e^{-3.043873} = 0.04765$  e  $e^{-2.737109} = 0.06475729$ . Estes valores são compatíveis com o número de observações de cada espécie, tendo sido avistados (para todas as espécies) 179 indivíduos no local D, menos de 10% dos 1860 indivíduos avistados no local C.

- (h) O modelo log-linear ajustado está associado à hipótese de independência, hipótese essa rejeitada pelo habitual teste  $\chi^2$  de independência (como seria de esperar, dado o problema em causa). Esta interpretação é corroborada pela proximidade do valor da soma de quadrados

dos resíduos de Pearson, calculada na alínea (e), com a soma de quadrados dos resíduos do desvio, que gera o desvio residual indicado nas listagens (sendo de esperar um valor próximo das duas somas de quadrados destas variantes de resíduos). No nosso caso, e como se pode verificar em cima, o desvio residual é  $D^* = 251\,164$ . Tendo em conta que o desvio residual se define à custa da diferença das log-verosimilhanças entre o modelo ajustado e o modelo saturado, e que o modelo saturado neste contexto corresponde a um modelo com efeitos de interação (e sem independência), pode calcular-se o  $p$ -value do desvio calculado  $D^* = 251\,164$  a partir duma distribuição  $\chi^2_{252}$  (uma vez que é essa a diferença entre o número de observações,  $n = 285$ , e o número de parâmetros do modelo,  $m = a + b - 1 = 33$ ). Uma vez que  $p \approx 0$ , rejeita-se a igualdade do modelo ajustado (associado à hipótese de independência) e do modelo saturado (sem qualquer restrição aos valores de  $\lambda_{ij}$ ). Ao optar pelo modelo completo, podemos afirmar que a hipótese de independência entre os factores locais e espécies, hipótese associada ao submodelo, deve ser rejeitada.

5. (a) Basta o comando `plot(Elisa1, pch=16)`.
- (b) Dada a natureza da variável resposta **emergencias**, que é uma variável de contagem do número de adultos que emergem na geração seguinte (não fazendo sentido impôr à partida um limite superior), pode admitir-se que essa variável resposta tenha distribuição de Poisson. O parâmetro  $\lambda$  duma Poisson é o seu valor esperado:  $E[Y] = \lambda$ , pelo que modelar  $\lambda$  significa modelar a tendência de fundo (valores médios) da variável resposta  $Y$ .
- (c) Essa função de ligação canónica da distribuição Poisson é o logaritmo:  $g(\lambda) = \ln(\lambda)$ . Assim, utilizar essa função de ligação corresponde a modelar a relação entre o preditor  $x$  (número de larvas de mosquito presentes no substrato, **esciarídeos**) e o logaritmo do número esperado de adultos na geração seguinte através da equação:

$$\ln(\lambda(x)) = \beta_0 + \beta_1 x \quad \Leftrightarrow \quad \lambda(x) = e^{\beta_0 + \beta_1 x}.$$

O gráfico indica como admissível a relação de tipo exponencial indicada, uma vez que uma curva exponencial (crescente) acompanha o padrão da nuvem de pontos.

- (d) O ajustamento do modelo obtém-se da seguinte forma e com os resultados indicados:

```
> Elisa1.glm <- glm(emergencias ~ esciarideos , family=poisson, data=Elisa1)
> summary(Elisa1.glm)
[...]
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2.632e+00	5.076e-02	51.85	<2e-16 ***
esciarideos	5.248e-04	3.209e-05	16.36	<2e-16 ***

---

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 513.00 on 56 degrees of freedom

Residual deviance: 244.22 on 55 degrees of freedom

AIC: 526.32

Number of Fisher Scoring iterations: 4

A curva ajustada tem assim equação  $y = e^{2.632+0.0005248x}$ . A interpretação do parâmetro  $\beta_1$  neste tipo de modelos pode fazer-se reparando que, para qualquer valor do preditor  $x$ , o valor esperado de  $Y$  é dado por  $\lambda(x) = e^{\beta_0} e^{\beta_1 x}$ . Aumentando o preditor numa unidade, o valor esperado de  $Y$  vem:  $\lambda(x+1) = e^{\beta_0} e^{\beta_1(x+1)} = e^{\beta_0} e^{\beta_1 x} e^{\beta_1}$ . Assim, tem-se

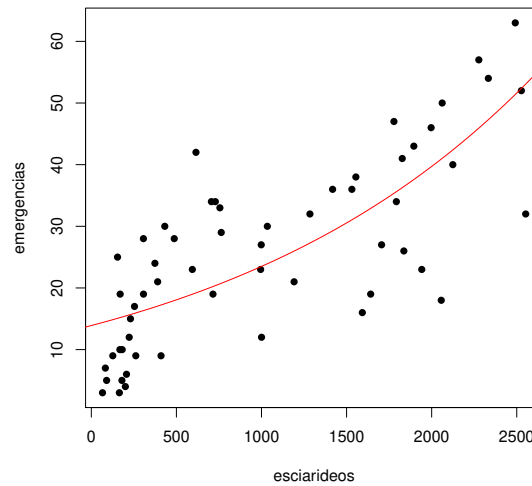
$$e^{\beta_1} = \frac{\lambda(x+1)}{\lambda(x)} \quad \Leftrightarrow \quad \beta_1 = \ln(\lambda(x+1)) - \ln(\lambda(x)).$$



Assim, cada aumento numa unidade no preditor  $x$  corresponde, em média, a multiplicar o valor esperado da variável resposta por  $e^{\beta_1}$ . No nosso exemplo, o valor estimado de  $\beta_1$  é  $b_1 = 0.0005248$ , pelo que aumentar numa unidade o número de larvas de mosquito presentes no substrato corresponde, em média, a multiplicar o número de adultos na geração seguinte por  $e^{0.0005248} = 1.000525$ . Este valor muito pequeno é também reflexo da escala de valores do preditor  $x$  (*esciarideos*), que varia entre valores próximos de zero e valores maiores do que 2500. Uma interpretação mais adequada a esta gama de valores observados poderia dizer respeito a um múltiplo de  $\beta_1$ . Por exemplo, através dum raciocínio análogo ao indicado na alínea 5d, pode concluir-se que a cada aumento de 100 unidades no valor do preditor, corresponde um factor multiplicativo de  $e^{0.0005248 \times 100} = 1.053881$  na razão  $\frac{\lambda(x+100)}{\lambda(x)}$ , pelo que o número médio de adultos emergentes aumentará, em média, cerca de 5% por cada 100 larvas de mosquito adicionais no substrato.

- (e) A curva ajustada acima pode ser traçada sobre a nuvem de pontos com o comando:

```
> curve(exp(2.632 + 0.0005248*x), from=-50, to=2700, col="red" , add=TRUE)
```



- (f) Os ICs (assintóticos) a 95% de confiança pedidos para cada  $\beta_j$  ( $j=1,2$ ) são da forma:

$$\left[ b_j - z_{0.025} \cdot \hat{\sigma}_{\hat{\beta}_j}, b_j + z_{0.025} \cdot \hat{\sigma}_{\hat{\beta}_j} \right],$$

onde  $b_j$  indica a estimativa do respectivo parâmetro  $\beta_j$ ,  $z_{0.025}$  é o quantil de ordem  $1-0.025=0.975$  numa distribuição  $\mathcal{N}(0,1)$ , e  $\hat{\sigma}_{\hat{\beta}_j}$  é o desvio padrão (assintótico) associado ao estimador  $\hat{\beta}_j$ , e dado pelo correspondente elemento diagonal da inversa da matriz de informação de Fisher,  $\mathbf{I}^{-1}$ . Ora,  $z_{0.025} = 1.959964$ . Os restantes valores são disponibilizados na listagem produzida pelo comando `summary(Elisa1.glm)`. Assim, e para o IC correspondente a  $\beta_1$ , tem-se  $b_1 = 0.0005248$  e  $\hat{\sigma}_{\hat{\beta}_1} = 0.00003209$ . O correspondente IC assintótico a 95% é  $]0.0004619, 0.0005877[$ . Este IC contém apenas valores positivos, pelo que o facto multiplicativo  $e^{\beta_1}$  que transforma  $\lambda(x)$  em  $\lambda(x+1)$  é, a 95% de confiança, maior do que um. Assim, é legítimo afirmar que a aumentos no número de mosquitos no substrato correspondem aumentos (em média) no número de adultos emergentes. (e é também legítimo afirmar, com 95% de confiança, que esse factor está entre  $e^{0.0004619} = 1.000462$  e  $e^{0.0005877} = 1.000588$ ).

---

Um IC para  $\beta_0$  obtém-se de forma análoga. No R, estes ICs são obtidos com o comando `confint.default`:

```
> confint.default(Elisa1.glm)
2.5 %      97.5 %
(Intercept) 2.5322823635 2.7312645384
esciarideos 0.0004619407 0.0005877286
```

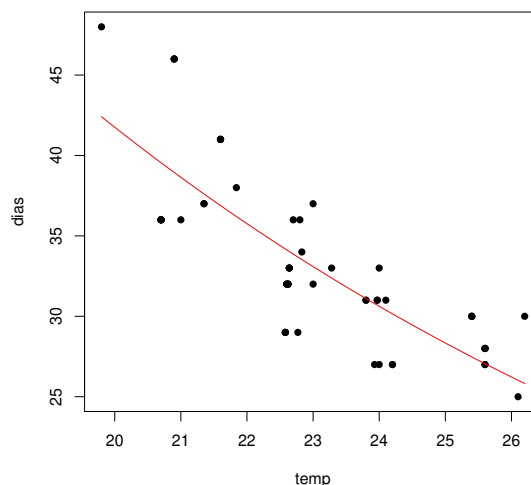
6. Pretende-se neste exercício modelar o número de dias entre a postura de ovos e a emergência de novos adultos (variável `dias`) com o preditor temperatura do meio ambiente (variável `temp`). Os dados estão na *data frame* `Elisa2`.

- (a) Vamos admitir que a componente aleatória  $Y$  (variável `dias`) é uma variável aleatória de contagem, com distribuição Poisson de parâmetro  $\lambda = E[Y]$  que depende do preditor  $x$  (variável `temp`) através da ligação canónica  $g(\lambda) = \ln(\lambda)$ . Assim, a equação de base do modelo é  $\ln(\lambda(x)) = \beta_0 + \beta_1 x$ , ou seja,  $\lambda(x) = e^{\beta_0 + \beta_1 x}$ . Esta relação de fundo exponencial é compatível com a nuvem de pontos dada no enunciado, que poderá ter a forma duma exponencial decrescente (a adequação, ou não, desta relação terá de ser confirmada com o ajustamento, e pode ainda ser melhor avaliada construindo uma nuvem de pontos entre `log(dias)` e `temp`, cuja tendência de fundo linear confirmará a adequação da relação antes descrita). Uma exponencial decrescente corresponde, na equação do modelo, a um valor negativo do parâmetro  $\beta_1$ . Eis o ajustamento no R deste modelo log-linear:

```
> Elisa2.glm1 <- glm(dias ~ temp, family=poisson, data=Elisa2)
> summary(Elisa2.glm1)
[...]
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.28241     0.32058  16.478 < 2e-16 ***
temp        -0.07753     0.01402  -5.528 3.23e-08 ***
---
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 47.194  on 56  degrees of freedom
Residual deviance: 16.203  on 55  degrees of freedom
AIC: 324.44
Number of Fisher Scoring iterations: 4
```

A diminuição considerável do desvio residual em relação ao desvio do modelo nulo indica que o preditor `temp` desempenha um papel importante na explicação da variabilidade no número de dias até à emergência, o que corresponde à indicação da nuvem de pontos. O valor estimado  $b_1 = -0.07753$  é negativo, indicando como foi acima referido, existir uma relação de tipo decrescente entre número de dias até à emergência e temperatura. O valor de prova (*p-value*) no teste (assintótico) a  $\beta_1 = 0$  é muito baixo ( $p = 3.23 \times 10^{-8}$ ), pelo que se rejeita a hipótese de  $\beta_1 = 0$ , havendo evidência estatisticamente significativa duma relação decrescente. A curva ajustada tem equação  $y = e^{5.28241 - 0.07753x}$ , e está indicada na nuvem de pontos em baixo, obtida com os seguintes comandos do R.

```
> plot(Elisa2, pch=16)
> curve(exp(5.28241-0.07753*x) , add=TRUE, col="red")
```



- (b) Nesta alínea pede-se para fazer uma única modificação ao modelo ajustado na alínea anterior: considerar que a componente aleatória  $Y$  (variável **dias**) tem distribuição Normal. Não se pede para alterar a função de ligação, que deve continuar a ser a ligação *log*. Assim, a equação do modelo mantém-se igual à da alínea anterior (podendo apenas trocar-se a letra  $\lambda$ , que corresponde à média numa distribuição  $Po(\lambda)$  por  $\mu$ , que a forma mais usual de indicar a média duma Normal). Por outras palavras, a equação do modelo é  $\mu(x) = e^{\beta_0 + \beta_1 x}$ . Não se trata dum modelo linear entre  $x$  e  $y$ , dada a transformação exponencial na relação entre  $E[Y]$  e  $x$ . Para ajustar este modelo no R, será necessário efectuar duas alterações no comando da alínea anterior: (i) especificar a família **gaussian** (palavra reservada no argumento **family** do comando **glm** para indicar uma componente aleatória Normal); e (ii) especificar explicitamente a função de ligação, uma vez que para a distribuição Normal a função de ligação canónica é a identidade, e não a função logaritmo. Felizmente que existe código escrito no R permitindo o ajustamento dum modelo com componente aleatória Normal e função de ligação logarítmica. Eis o ajustamento e os respectivos resultados:

```
> Elisa2.glm2 <- glm(dias ~ temp, family=gaussian(link=log), data=Elisa2)
> summary(Elisa2.glm2)
[...]
```

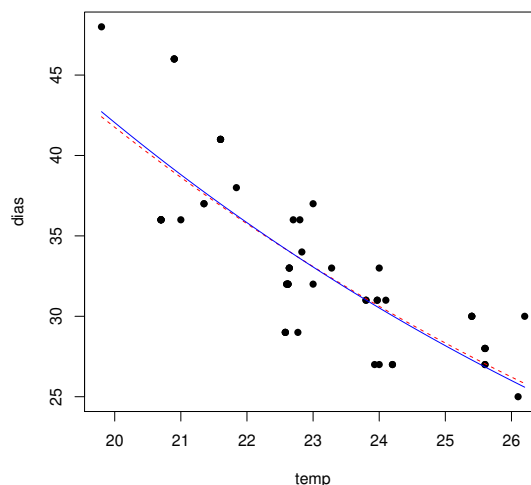
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.340326	0.181465	29.429	< 2e-16 ***
temp	-0.080079	0.008015	-9.992	5.75e-14 ***

```
---
(Dispersion parameter for gaussian family taken to be 10.4694)
Null deviance: 1634.67 on 56 degrees of freedom
Residual deviance: 575.82 on 55 degrees of freedom
AIC: 299.58
Number of Fisher Scoring iterations: 4
```

A curva ajustada tem agora a equação  $y = e^{5.340326 - 0.080079x}$  e pode ser traçada sobre a nuvem de pontos com os seguintes comandos (tendo sido deixada, a tracejada, a curva obtida a partir do modelo log-linear ajustado na alínea anterior).

```
> plot(Elisa2, pch=16)
> curve(exp(5.28241-0.07753*x) , add=TRUE, col="red", lty="dashed")
> curve(exp(5.340326-0.080079*x) , add=TRUE, col="blue")
```



Como se pode comprovar, as curvas ajustadas nos dois modelos são diferentes, uma vez que as estimativas  $b_0$  e  $b_1$  são diferentes (pois as verosimilhanças que foram maximizadas são diferentes para componentes aleatórias Normal ou Poisson). No entanto, as curvas ajustadas são semelhantes, o que é também natural e dá alguma tranquilidade relativamente à robustez do ajustamento.

Finalmente, considere-se o valor do desvio, em relação ao qual a afirmação feita no enunciado não é defensável. Não é um valor directamente comparável com o valor do desvio residual obtido no modelo Poisson, uma vez que (i) as funções de verosimilhança que estão na base da definição do desvio são diferentes; e (ii) em modelos com componente aleatória Normal o Desvio e o Desvio reduzido não coincidem, uma vez que existe parâmetro de dispersão  $\phi = \sigma^2$  desconhecido. Neste exemplo, o parâmetro de dispersão estimado, se fôr admitido comum a todas as observações, é dado na listagem:  $\hat{\phi} = 10.4694$ . Recorde-se que esta estimativa é a soma dos quadrados dos resíduos de Pearson, a dividir pelo número de observações ( $n=57$ ), menos o número de parâmetros do modelo ( $m=p+1=2$ ), isto é

```
> sum(residuals(Elisa2.glm2, type="pearson")^2)/55
[1] 10.46937
```

- (c) Nesta alínea pede-se para ajustar um modelo linear que corresponda, o melhor possível, aos modelos acima ajustados. Embora a relação entre `dias` e `temp` evidenciada na nuvem de pontos do enunciado pareça razoavelmente linear, o modelo linear que melhor corresponde aos anteriores será o modelo resultante da linearização da relação exponencial que, como foi visto na primeira parte da matéria deste módulo, é a relação linear entre o *logaritmo* de  $y$  e  $x$ , ou seja,  $\ln(y) = \beta_0 + \beta_1 x$ , a que corresponde a equação  $y = e^{\beta_0 + \beta_1 x}$ . Eis o ajustamento obtido no R:

```
> Elisa2.lm <- lm(log(dias) ~ temp, data=Elisa2)
> summary(Elisa2.lm)
[...]
```

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.203766	0.167995	30.98	< 2e-16 ***
temp	-0.074275	0.007281	-10.20	2.72e-14 ***

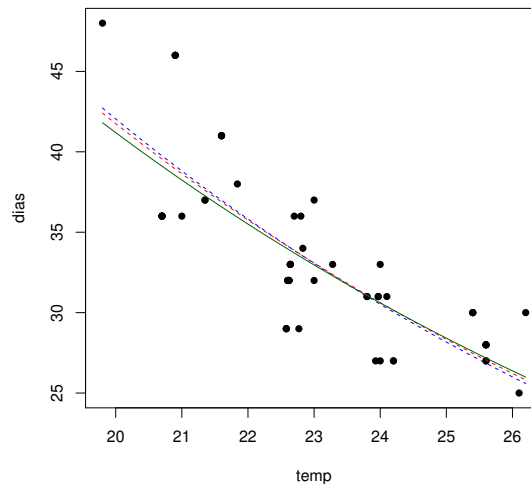
```
---
Residual standard error: 0.09165 on 55 degrees of freedom
```

---

```
Multiple R-squared: 0.6542, Adjusted R-squared: 0.6479
F-statistic: 104.1 on 1 and 55 DF, p-value: 2.72e-14
```

A curva ajustada tem agora equação  $y = e^{5.203766 - 0.074275x}$ . Eis o seu ajustamento, obtido com os comandos:

```
> plot(Elisa2, pch=16)
> curve(exp(5.28241-0.07753*x) , add=TRUE, col="red", lty="dashed")
> curve(exp(5.340326-0.080079*x) , add=TRUE, col="blue", lty="dashed")
> curve(exp(5.203766-0.074275*x) , add=TRUE, col="darkgreen")
```



Como se pode verificar, os três modelos ajustados neste caso produzem curvas semelhantes. Em termos de modelo linear, este último modelo explica cerca de 65,4% da variabilidade observada *nos log-dias*. Mas, tratando-se dum modelo linear, é um caso particular dum modelo linear generalizado, e é legítima a pergunta sobre qual o seu desvio residual. Para obter essa informação, podemos re-ajustar o mesmo modelo, usando agora o comando `glm`. Agora, será necessário substituir a função de ligação logarítmica pela transformação logarítmica da variável resposta `dias`, e especificar a função de ligação *identidade*. Sendo esta a função de ligação canónica para componentes aleatórias Normais, não será necessário especificá-la. Eis o ajustamento:

```
> Elisa2.glm3 <- glm(log(dias) ~ temp, family=gaussian, data=Elisa2)
> summary(Elisa2.glm3)
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.203766   0.167995   30.98 < 2e-16 ***
temp        -0.074275   0.007281  -10.20 2.72e-14 ***
---
(Dispersion parameter for gaussian family taken to be 0.008400088)
Null deviance: 1.3362 on 56 degrees of freedom
Residual deviance: 0.4620 on 55 degrees of freedom
AIC: -106.71
Number of Fisher Scoring iterations: 2
```

Como se pode confirmar, a curva ajustada é exactamente a mesma que a obtida com o modelo linear após a linearização da relação. O valor de desvio residual ( $D=0.4620$ ) não é

igual ao obtido no modelo da alínea anterior, uma vez que a distribuição Normal diz agora respeito à transformação logarítmica de  $Y$ , e não à própria variável  $Y$ . O valor estimado do parâmetro de dispersão  $\phi$  é agora  $\hat{\phi} = 0.008400088$  e corresponde ao Quadrado Médio Residual do modelo linearizado (como se viu nas aulas). De facto, e recordando que  $QMRE$  é o quadrado do valor indicado por **Residual standard error** na listagem de resultados dum modelo linear, temos:  $0.09165^2 = 0.008399722$  que, a menos de erros de arredondamento, corresponde ao valor indicado de  $\hat{\phi}$ .

7. (a) Substituindo, na expressão original da função densidade Gama,  $\nu$  por  $\alpha$  e  $\frac{\mu}{\nu}$  por  $\beta$ , tem-se:

$$\frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} e^{-\frac{\nu y}{\mu}} = \frac{1}{\left(\frac{\mu}{\nu}\right)^\nu \Gamma(\nu)} y^{\nu-1} e^{-\frac{y}{\left(\frac{\mu}{\nu}\right)}} = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}},$$

como se queria mostrar. Nesta nova parametrização (que foi usada pela Prof. Manuela Neves nas aulas iniciais desta disciplina) tem-se  $E[Y] = \alpha\beta$  e  $V[Y] = \alpha\beta^2$ .

- (b) É imediato a partir da expressão anterior, tomando  $n = \alpha$  e  $\gamma = \frac{1}{\beta}$ . Tem-se  $E[Y] = \frac{n}{\gamma}$  e  $V[Y] = \frac{n}{\gamma^2}$ .

8. (a) Num Modelo Linear Generalizado, designa-se função de ligação a uma função invertível  $g$ , com boas propriedades de regularidade, que relaciona o valor esperado da variável resposta  $Y$  com a combinação linear dos preditores, através da seguinte equação:

$$g(E[Y]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

- (b) Num MLG, designa-se resíduo do desvio da observação  $i$  com que se ajustou o modelo, à raiz quadrada da contribuição dessa  $i$ -ésima observação para o valor do desvio. O sinal da raiz quadrada é atribuído tendo em conta a relação entre o valor da variável resposta nessa observação  $i$  (i.e.,  $y_i$ ) e o valor médio estimado para uma observação com os mesmos valores dos preditores que a observação  $i$  (isto é,  $\hat{\mu}_i$ ). Se  $y_i \geq \hat{\mu}_i$  associa-se o sinal positivo ao resíduo, caso contrário o sinal negativo. Assim, sendo o desvio  $D = \sum_{i=1}^n d_i$ , tem-se como resíduo do desvio da observação  $i$ ,  $r_i^D = \text{sin}(\text{al}(y_i - \hat{\mu}_i))\sqrt{d_i}$ .

9. (a) i. Foi ajustado um modelo linear generalizado em que a componente aleatória (variável resposta) é uma variável dicotómica com distribuição de Bernoulli, sendo o valor “1” associado à espécie **carduorum** dos escaravelhos, e o valor “0” à espécie **oleracea**; a componente sistemática é definida pela combinação linear de quatro preditores (todos numéricos), designados **TG**, **Elytra**, **Second.Antenna** e **Third.Antenna** (descritos no enunciado); e finalmente a função ligação é (como é característico duma regressão logística) a função de ligação canónica para variáveis com distribuição Bernoulli, ou seja a função *logit*:  $g(p) = \ln\left(\frac{p}{1-p}\right)$ , sendo  $p$  quer o valor esperado de  $Y$ ,  $E[Y]$ , quer a probabilidade de um êxito (para os valores observados dos preditores). Assim, a probabilidade dum escaravelho observado ser da espécie **Haltica carduorum**, dados os valores dos preditores é dado por

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 TG + \beta_2 Elytra + \beta_3 Second.Antenna + \beta_4 Third.Antenna$$

$$\Leftrightarrow p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 TG + \beta_2 Elytra + \beta_3 Second.Antenna + \beta_4 Third.Antenna)}}$$

- 
- ii. O desvio é praticamente nulo ( $D = 4.7616 \times 10^{-10}$ ), sendo o desvio do modelo nulo  $D = 54.04$ . Recorde-se que o conceito de desvio envolve a diferença da log-verosimilhança do modelo ajustado com a log-verosimilhança dum modelo saturado. Neste caso, o modelo ajustado tem uma log-verosimilhança essencialmente idêntica à de um modelo saturado, mas não por ser efectivamente um modelo saturado. Não se trata dum problema de sobreparametrização, uma vez que existem  $n = 39$  observações, e apenas  $p + 1 = 5$  parâmetros no modelo. Trata-se apenas dum muito bom modelo discriminante, que permite uma discriminação quase perfeita para as  $n = 39$  observações efectuadas. Em certo sentido, é como se se tratasse duma regressão linear com  $R^2 = 1$ , mas não porque haja um número de observações igual ao número de parâmetros do modelo.
  - iii. O valor estimado para  $\beta_3$  é  $b_3 = 7.634$ . Como se viu (acetato 48), tal significa que para cada micrómetro adicional no comprimento do segundo segmento antenal, a razão de probabilidades  $\frac{p}{1-p}$  aumenta num factor de  $e^{7.634} = 2067.303$  vezes.
  - iv. Com base na listagem de resultados dada no enunciado, qualquer dos quatro preditores pode ser excluído, sem perda de qualidade relevante. Os valores de prova (*p-values*) correspondentes aos testes a  $\beta_j = 0$  são todos indistinguíveis da unidade! Mesmo assim, e com base nos valores das estatísticas  $z$ , a excluir uma variável seria aquela com o valor de  $z_{calc}$  mais próximo de zero, ou seja, o preditor **Third.Antenna**.
- (b) O modelo final obtido excluiu o preditor comprimento do terceiro segmento de antena. O desvio ficou praticamente igual ( $D = 3.846 \times 10^{-10}$ ), o que é coerente com a resposta dada na alínea anterior. Note-se que a equação para  $p$  resultante deste modelo ajustado alterou substancialmente as estimativas dos coeficientes dos preditores restantes. Note-se ainda que a exclusão de um segundo preditor (que seria o comprimento do segundo segmento de antena) já gera um desvio diferente de zero (mais concretamente  $D = 9.8414$ ), o que significaria que o ajustamento deste modelo já não seria igual ao de um modelo saturado.
- (c) Considera-se agora o submodelo cujos preditores são apenas **TG** e **Elytra**.
- i. Pretende-se inicialmente testar se este submodelo difere do modelo inicial, com os quatro preditores.
- Hipóteses:**  $H_0 : \beta_3 = \beta_4 = 0$  vs.  $H_1 : \beta_3 \neq 0 \vee \beta_4 \neq 0$ .
- Estatística do teste:**  $\Lambda = D_s^* - D_c^* \sim \chi_2^2$ , onde  $D^*$  indica o desvio, e  $s$  e  $c$  indicam submodelo e modelo completo, respectivamente. No caso duma componente aleatória com distribuição de Bernoulli, desvio e desvio reduzido são a mesma coisa, como se viu nas aulas. Os graus de liberdade (2) resultam do facto de ser essa a diferença entre o número de parâmetros do modelo completo (5) e do submodelo (3).
- Nível de significância:**  $\alpha = 0.05$ .
- Região Crítica (Unilateral direita):** Rejeitar  $H_0$  se  $\Lambda_{calc} > \chi_{\alpha(2)}^2 \approx 5.991465$ .
- Conclusões:** O valor calculado da estatística do teste é  $\Lambda_{calc} = 9.8414 - 0 = 9.8414 > 5.991465$ , pelo que se rejeita a hipótese nula, ou seja, admite-se que o modelo e o submodelo diferem significativamente, ao nível  $\alpha = 0.05$ . O valor de prova (*p-value*) associado ao valor calculado da estatística é  $p = 0.0073$ , como se pode verificar:
- ```
> 1-pchisq(9.8414,2)
[1] 0.007294023
```
- No entanto, convém ter presente que a distribuição da estatística do teste de Wilks é apenas assintótica e a dimensão da amostra ( $n = 39$ ) não é muito grande, pelo que haverá que ser cauteloso na consideração deste teste.

- ii. A probabilidade de pertença à espécie *carduorum* é dada, no modelo ajustado, por:

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + b_1 TG + b_2 Elytra)}} = \frac{1}{1 + e^{-(10.1559 - 0.4271 TG + 0.2505 Elytra)}} .$$

Assim, para um escaravelho com  $TG = 200$  e  $Elytra = 250$ , a probabilidade de ser da espécie *carduorum* é estimada por:

$$\hat{p} = \frac{1}{1 + e^{-(10.1559 - 0.4271 \times 200 + 0.2505 \times 250)}} = 3.242703 \times 10^{-6} .$$

A probabilidade de ser da espécie *oleracea* será a diferença para a unidade:

$$1 - \hat{p} = 1 - 0.000003243 = 0.999996757 .$$

Assim, é quase seguro que um escaravelho com as referidas características seja da espécie *oleracea*. Esta conclusão é sustentada por uma análise dos dados, já que os escaravelhos da espécie *carduorum* têm, em geral, valores de  $TG$  menores que 200 e valores de  $Elytra$  maiores que 250.

- (d) O modelo considerado nesta alínea utiliza a função de ligação log-log do complementar:

$$\hat{p} = 1 - e^{-e^{b_0 + b_1 TG + b_2 Elytra}} = 1 - e^{-e^{7.78272 - 0.33889 TG + 0.19769 Elytra}} .$$

- i. O facto de os pontos estarem muito próximos da bissetriz de equação  $y = x$  indica que as probabilidades estimadas de pertença à espécie *carduorum*, obtidas nos dois modelos, são muito semelhantes. Isso é igualmente verdade para a “observação solitária” 19. Embora as probabilidades de pertença à referida espécie sejam próximas de 0.5 (ou seja, trata-se dum escaravelho difícil de classificar através destes modelos), essa probabilidade estimada é muito semelhante nos dois modelos, como se depreende da proximidade do ponto à recta bissetriz.
- ii. Com base na informação disponível, o modelo desta alínea será preferível, uma vez que tem um desvio (logo também um AIC) mais baixo. Assinale-se que os valores dos desvios são comparáveis, uma vez que se trata de modelos com a mesma componente aleatória e ajustados com base nos mesmos dados. A função de ligação log-log do complementar parece ser mais adequada à natureza da relação entre a probabilidade de pertença à espécie *carduorum* e a componente sistemática do modelo.
- iii. A probabilidade de pertença à espécie *carduorum* é agora dada por

$$\hat{p} = 1 - e^{-e^{7.78272 - 0.33889 \times 200 + 0.19769 \times 250}} = 2.560321 \times 10^{-5} .$$

Tal como para o modelo de regressão logística considerado na alínea anterior, trata-se duma probabilidade muito próxima de zero. Também com este modelo, um escaravelho com os valores referidos de  $TG$  e  $Elytra$  seria classificado como sendo da espécie *oleracea*.

10. Ver acetatos das aulas.

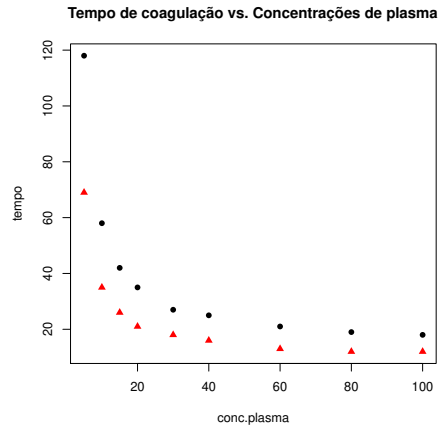
11. Os dados encontram-se na *data frame* `sangue`.

- (a) O comando `R` para construir o gráfico pedido pode ser o seguinte:

```
> plot(tempo ~ conc.plasma, col=as.numeric(lote), pch=as.numeric(lote)+15,
       data=sangue, main="Tempo de coagulação vs. Concentrações de plasma")
```



A utilização do comando `as.numeric` no argumento `pch` visa converter os níveis do factor `lote` para números naturais (neste caso, 1 e 2). A parte “+15” do argumento `pch` visa garantir a utilização dos símbolos 16 e 17, que correspondem respectivamente aos círculos e triângulos preenchidos. Eis o resultado:

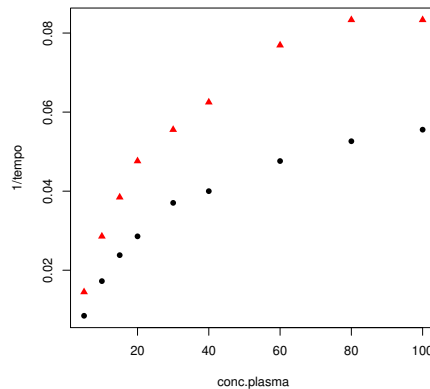


A forma da relação sugere uma curva de tipo hiperbólico, ou eventualmente uma exponencial decrescente.

- (b) Uma relação do tipo  $y = \frac{1}{\beta_0 + \beta_1 x}$  corresponde a uma relação linear entre  $y^* = \frac{1}{y}$  e  $x$ . Assim, para validar se a relação de tipo hiperbólico proposta é adequada, pode construir-se uma nuvem de pontos relacionando os recíprocos de  $y$  com  $x$ :

```
> plot(1/tempo ~ conc.plasma, col=as.numeric(lote), pch=as.numeric(lote)+15,
      data=sangue)
```

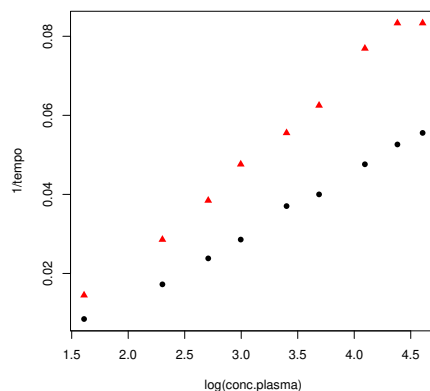
O gráfico resultante é:



Uma vez que a relação obtida não é linear, deve concluir-se que a relação de tipo  $tempo = \frac{1}{\beta_0 + \beta_1 conc}$  que foi proposta não é a mais adequada.

- (c) Efectuando uma transformação logarítmica da variável preditora, a relação entre  $y^* = \frac{1}{y}$  e  $x^* = \ln(x)$  já é bem aproximada por uma relação linear:

```
> plot(1/tempo ~ log(conc.plasma), col=as.numeric(lote), pch=as.numeric(lote)+15,
      data=sangue)
```



Assim, a equação  $y = \frac{1}{\beta_0 + \beta_1 \ln(x)}$  parece ser uma descrição adequada para relacionar tempo de coagulação ( $y$ ) e concentração de plasma ( $x$ ).

(d) São pedidos dois modelos lineares generalizados, ambos com o preditor  $X = \ln(\text{conc})$  e variável resposta  $Y = \text{tempo}$ , e ambos com função de ligação recíproco:  $g(\mu) = \frac{1}{\mu}$ .

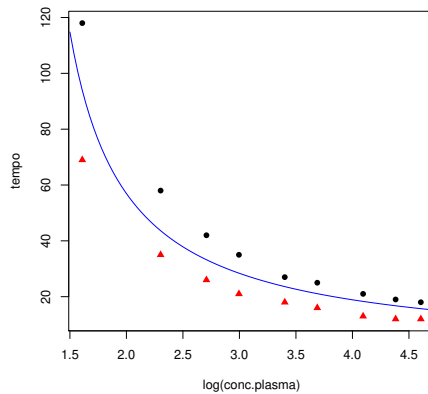
i. No primeiro caso, sugere-se admitir que  $Y$  tem distribuição Normal. O comando R para ajustar este modelo é:

```
> sangue.glmN <- glm(tempo~log(conc.plasma),family=gaussian(link="inverse"),
+                               data=sangue)
> sangue.glmN
Call:
glm(formula=tempo~log(conc.plasma), family=gaussian(link="inverse"), data=sangue)
Coefficients:
      (Intercept)    log(conc.plasma)
      -0.01784         0.01770
Degrees of Freedom: 17 Total (i.e. Null); 16 Residual
Null Deviance:      11880
Residual Deviance: 1875  AIC: 140.7
```

Assim, a curva ajustada é  $y = \frac{1}{-0.01784+0.01770x}$ , curva essa que é sobreposta à nuvem de pontos de  $Y = \text{tempo}$  vs.  $X = \ln(\text{conc})$ , através dos comandos

```
> plot(tempo~log(conc.plasma), col=as.numeric(lote), pch=as.numeric(lote)+15,
+                               data=sangue)
> curve(1/(-0.01784+0.01770*x), from=1.5, to=5, add=TRUE, col="blue")
```

produzindo o seguinte gráfico:



- ii. O ajustamento dum modelo com variável resposta gama obtém-se de forma análoga, embora neste caso (como é dito no enunciado) não seja necessário especificar a função de ligação usada, uma vez que a ligação canónica para distribuição gama é precisamente a função recíproco. Assim, o comando necessário e o respectivo resultado são:

```
> sangue.glmG <- glm(tempo ~ log(conc.plasma), family=Gamma, data=sangue)
> sangue.glmG
Call: glm(formula=tempo ~ log(conc.plasma), family=Gamma, data=sangue)
Coefficients:
```

```
      (Intercept)    log(conc.plasma)
      -0.01963      0.01861
```

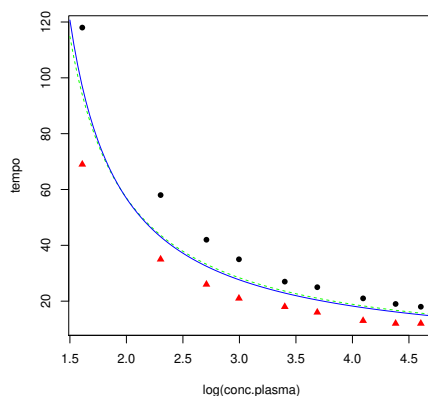
```
Degrees of Freedom: 17 Total (i.e. Null); 16 Residual
```

```
Null Deviance:      7.709
```

```
Residual Deviance: 1.018 AIC: 123.2
```

A curva ajustada é indicada em baixo (a azul), comparada com a curva (a verde a tracejado) obtida com o modelo Normal da alínea anterior. As duas curvas são quase indistinguíveis. O gráfico foi obtido com os seguintes comandos:

```
> plot(tempo~log(conc.plasma),col=as.numeric(lote),pch=as.numeric(lote)+15,data=sangue)
> curve(1/(-0.01784+0.01770*x), from=1.5, to=5, add=TRUE, col="green", lty="dashed")
> curve(1/(-0.01963+0.01861*x), from=1.5, to=5, add=TRUE, col="blue")
```



- (e) Não é possível uma comparação com base nos desvios de cada modelo, uma vez que as distribuições associadas à componente aleatória são diferentes em cada caso (pelo que os valores das log-verosimilhanças não são comparáveis). É possível efectuar testes comparando cada modelo com o respectivo modelo nulo, mas os resultados são praticamente idênticos

---

(como seria de esperar, dada a grande proximidade nas curvas ajustadas):

```
> anova(sangue.glmN, test="LRT")
```

Analysis of Deviance Table

Model: gaussian, link: inverse

Response: tempo

Terms added sequentially (first to last)

|                  | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi)      |
|------------------|----|----------|-----------|------------|---------------|
| NULL             |    |          | 17        | 11884.5    |               |
| log(conc.plasma) | 1  | 10009    | 16        | 1875.4     | < 2.2e-16 *** |

---

```
> anova(sangue.glmG, test="LRT")
```

Analysis of Deviance Table

Model: Gamma, link: inverse

Response: tempo

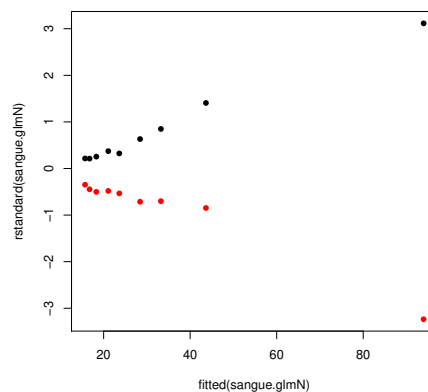
Terms added sequentially (first to last)

|                  | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi)      |
|------------------|----|----------|-----------|------------|---------------|
| NULL             |    |          | 17        | 7.7087     |               |
| log(conc.plasma) | 1  | 6.6904   | 16        | 1.0183     | < 2.2e-16 *** |

Em ambos os casos a conclusão é que os modelos ajustados são claramente melhores do que os respectivos modelos nulos, e em ambos os casos com *p-values* indistinguíveis de zero.

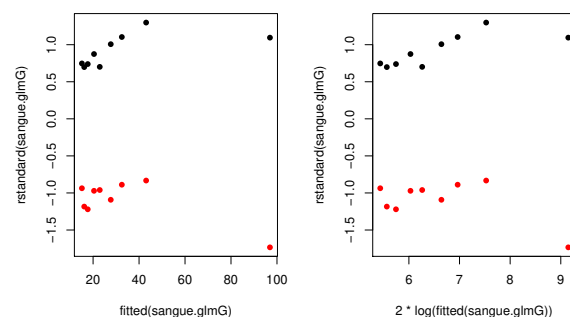
Consideremos o gráfico dos resíduos (estandardizados) *vs.* os valores médios ajustados pelo modelo, reproduzidos na seguinte nuvem de pontos, onde se utilizaram cores diferentes para distinguir o lote de proveniência das observações. O gráfico foi produzido com o comando:

```
> plot(fitted(sangue.glmN), rstandard(sangue.glmN), col=as.numeric(sangue$lote), pch=16)
```



Como se pode constatar, existe um padrão nos gráficos que distingue as observações de cada lote, sendo as observações do lote 1 indicadas a preto. Esse padrão reflecte o facto de não ter sido utilizada a informação relativa ao lote de proveniência no modelo ajustado. Pelo que se pode ver no gráfico, esse factor tem alguma importância, uma vez que as observações do Lote 1 são em geral maiores (têm os resíduos positivos), facto que se pode também observar na tabela dos dados ou na nuvem de pontos com a curva ajustada sobreposta (ver acima). Assim, este padrão sugere que o modelo deve prever *efeitos de lote*, algo que será estudado

nas alíneas seguintes. Por outro lado, existe um segundo padrão visível neste gráfico: a forma em funil dos resíduos, que sugere que os resíduos (standardizados) crescem em valor absoluto com o aumento do valor de  $\hat{\mu}$ . Assim, a dispersão dos valores da área foliar não parece ser constante. Haveria duas formas de torcear este problema: manter as características do modelo ajustado (distribuição Normal, função de ligação logarítmica), mas prever parâmetros de dispersão diferenciados para as diferentes observações (matéria que não foi dada nesta disciplina), ou alternativamente, passar para a distribuição Gama da variável resposta, uma vez que nessa distribuição (e mesmo admitindo parâmetro de dispersão  $\phi$  constante para todas as observações), a variância das observações será proporcional ao quadrado da média (acetato 141). Eis (à esquerda), o gráfico de resíduos standardizados *vs.* valor esperado ajustado, correspondente a admitir uma variável resposta com distribuição Gama e (à direita), um gráfico análogo com a transformação logarítmica no eixo horizontal (como sugerido no acetato 161), novamente distinguindo os lotes de proveniência das observações através da cor.



É visível que o “efeito funil” fica bastante mitigado, ou desaparece mesmo. Quanto à presença dum efeito de lote, será estudado nas alíneas seguintes.

(f) Os modelos pedidos ajustam-se no R através dos seguintes comandos. Para o modelo Normal:

```
> sangue.glmNlote <- glm(formula = tempo ~ log(conc.plasma)*lote,
+                          family=gaussian(link = "inverse"), data=sangue)
> sangue.glmGlote <- glm(formula=tempo ~ log(conc.plasma)*lote,
+                          family=Gamma(link = "inverse"), data=sangue)
```

Eis os resultados obtidos no ajustamento do modelo Normal:

```
> sangue.glmNlote
Call:
glm(formula=tempo~log(conc.plasma)*lote,family=gaussian(link="inverse"),data=sangue)
Coefficients:
      (Intercept)      log(conc.plasma)      lote2  log(conc.plasma):lote2
      -0.014903      0.014498      -0.007170      0.008181
Degrees of Freedom: 17 Total (i.e. Null); 14 Residual
Null Deviance: 11880
Residual Deviance: 34.49 AIC: 72.79
```

Numa primeira observação, boa parte do desvio associado ao modelo nulo desapareceu com este modelo. O desvio agora obtido é também bastante inferior ao obtido no modelo Normal, mas sem a consideração dos lotes (sendo esse desvio 1875.4, e agora apenas 34.49). Para o modelo Gama obtiveram-se os resultados:

```
> sangue.glmGloTe
Call: glm(formula=tempo~log(conc.plasma)*lote,family=Gamma(link="inverse"),data=sangue)
Coefficients:
      (Intercept)      log(conc.plasma)      lote2      log(conc.plasma):lote2
      -0.016554         0.015343        -0.007354         0.008256
Degrees of Freedom: 17 Total (i.e. Null); 14 Residual
Null Deviance:      7.709
Residual Deviance: 0.0294 AIC: 63.2
```

Tal como no modelo Normal, a introdução de efeitos de lote permitiu reduzir o desvio de 1.0183 (em comparação com 7.709 do modelo Normal nulo) para apenas 0.0294, uma redução considerável que é provavelmente significativa, o que adiante se estudará.

(g) No modelo Normal, a curva ajustada para as observações do lote 1 é

$$y = \frac{1}{-0.01490 + 0.014498x},$$

onde  $y$  indica tempo e  $x$  log-concentração de plasma. Para as observações do lote 2 tem-se:

$$y = \frac{1}{(-0.01490 - 0.007170) + (0.014498 + 0.008181)x} = \frac{1}{-0.02207 + 0.022679x}.$$

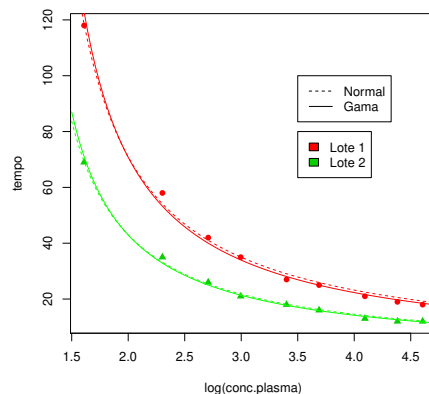
No modelo Gama, a curva ajustada para as observações do lote 1 é

$$y = \frac{1}{-0.016554 + 0.015343x},$$

onde  $y$  indica tempo e  $x$  log-concentração de plasma. A curva ajustada para as observações do lote 2 é

$$y = \frac{1}{(-0.016554 - 0.007354) + (0.015343 + 0.008256)x} = \frac{1}{-0.023908 + 0.023599x}.$$

Admitiu-se que existe uma relação linear entre  $tempo^{-1}$  e log-concentração, sendo que nessa relação linear o declive é maior no lote 2, e a ordenada na origem é menor (sendo ambas as ordenadas na origem negativas). Tendo em conta que, nesta relação linear,  $y$  é  $\frac{1}{tempo}$ , podemos afirmar que esta recíproco cresce mais depressa com a log-concentração no Lote 2, ou seja que *o tempo de coagulação* decresce mais depressa com a log-concentração no Lote 2, embora partindo dum tempo de coagulação mais baixo. Eis as curvas ajustadas pelos modelos Normal e Gama, mas prevendo efeitos de lote.



O gráfico e as legendas acima indicadas foram produzidos com os seguintes comandos do R.

```
> plot(tempo~log(conc.plasma),col=as.numeric(lote)+1,pch=as.numeric(lote)+15,data=sangue)
> legend(3.5,80, legend=paste("Lote",c(1,2)), fill=c(1,2)+1)
> legend(3.5,100, legend=c("Normal","Gama"), lty=c("dashed","solid"))
> curve(1/(-0.01490+0.014498*x), from=1.5, to=5, add=TRUE, col="red", lty="dashed")
> curve(1/(-0.016554+ 0.015343*x), from=1.5, to=5, add=TRUE, col="red")
> curve(1/(-0.023908+0.023599*x), from=1.5, to=5, add=TRUE, col="green")
> curve(1/(-0.02207+0.022679*x), from=1.5, to=5, add=TRUE, col="green", lty="dashed")
```

- (h) Existem duas questões que vale a pena abordar. A primeira é a de saber se a incorporação do factor `lote` no modelo melhorou significativamente o ajustamento. A resposta é (previsivelmente, após verificar o ajustamento das curvas obtidas à nuvem de pontos) positiva. Tal facto pode ser confirmado efectuando os testes de Wilks para comparar o modelo completo, cuja equação é

$$y = \frac{1}{(\beta_0 + \alpha_{0:j}) + (\beta_1 + \alpha_{1:j})x}$$

sendo  $\alpha_{i:j}$  o acréscimo ao parâmetro  $\beta_i$ , associado ao lote  $j$  (com  $\alpha_{0:1} = \alpha_{1:1} = 0$ ). com o submodelo, que prevê uma única relação para ambos os lotes:

$$y = \frac{1}{\beta_0 + \beta_1 x}.$$

Formalizando este teste, quer para os modelos baseados na Normal, quer para os modelos baseados na Gama:

**Hipóteses:**  $H_0 : \alpha_{0:2} = \alpha_{1:2} = 0$  vs.  $H_1 : \alpha_{0:2} \neq 0 \vee \alpha_{1:2} \neq 0$ .

**Estatística do teste:**  $\Lambda = D_s^* - D_c^* \sim \chi^2_2$ , sob  $H_0$ , onde  $D^*$  indica o desvio *reduzido*, e  $s$  e  $c$  indicam submodelo e modelo completo, respectivamente. Os graus de liberdade (2) resultam do facto de ser essa a diferença entre o número de parâmetros do modelo completo (4) e do submodelo (2).

**Nível de significância:**  $\alpha = 0.05$ . Note-se que a distribuição desta estatística de teste é apenas assintótica e o número de observações neste caso não é muito grande, pelo que haverá sempre alguma dúvida sobre a validade da conclusão.

**Região Crítica (Unilateral direita):** Rejeitar  $H_0$  se  $\Lambda_{calc} > \chi^2_{\alpha(2)} \approx 5.991465$ .

**Conclusões:** O valor calculado da estatística do teste tem de ter em conta o facto de os desvios apresentados nas listagens produzidas pelo R serem os desvios *não* reduzidos, enquanto que a estatística baseia-se nos desvios reduzidos. Recorde-se (acetato 115) que a relação entre os desvios  $D$  e os desvios reduzidos  $D^*$  é:  $D^* = \frac{D}{\phi}$ . Assim, para calcular a estatística do teste será necessário ter uma estimativa do parâmetro de dispersão  $\phi$  (que se admite comum para todas as observações) e depois usar essa estimativa,  $\hat{\phi}$ , no cálculo da estatística do teste:  $\Lambda_{calc} = D_s^* - D_c^* = \frac{D_s - D_c}{\hat{\phi}}$ . Para obter as estimativas do parâmetro de dispersão  $\phi$  precisamos do componente **dispersion** do comando **summary**. É conveniente utilizar as estimativas obtidas com os modelos completos, cujos ajustamento são necessariamente melhores. Para cada modelo, temos:

```
> summary(sangue.glmNlote)$disp
[1] 2.463579
> summary(sangue.glmGlote)$disp
[1] 0.002129707
```

Assim, e tendo em conta os valores dos desvios indicados nas alíneas anteriores, a estatística calculada para o teste de Wilks é, no caso dos modelos Normais,  $\Lambda_{calc} = \frac{1875.4 - 34.49}{2.463579} = 747.2502 \gg 5.991465$ . Para os modelos Gama, tem-se  $\Lambda_{calc} = \frac{1.0183 - 0.0294}{0.002129707} = 464.3362 \gg 5.991465$ . Os valores tão elevados de  $\Lambda_{calc}$  são, em ambos os casos, seguramente significativos mesmo para níveis de significância muito inferiores a  $\alpha = 0.05$ . Esta constatação é reconfortante, dadas as dúvidas (já referidas) associadas à natureza assintótica da distribuição da estatística do teste.

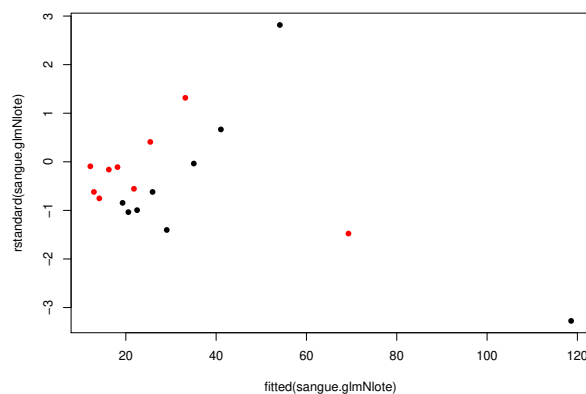
Os resultados dos testes são confirmados pela utilização do comando `anova` do R.

```
> anova(sangue.glmN, sangue.glmNlote, test="LRT")
Analysis of Deviance Table

Model 1: tempo ~ log(conc.plasma)
Model 2: tempo ~ log(conc.plasma) * lote
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         16      1875.38
2          14        34.49  2    1840.9 < 2.2e-16 ***
---
> anova(sangue.glmG, sangue.glmGlote, test="LRT")
Analysis of Deviance Table

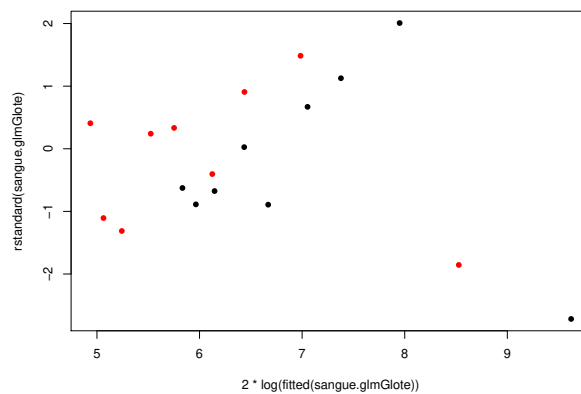
Model 1: tempo ~ log(conc.plasma)
Model 2: tempo ~ log(conc.plasma) * lote
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         16        1.0183
2          14        0.0294  2    0.98886 < 2.2e-16 ***
```

O segundo aspecto para analisar diz respeito à adequação dos modelos com efeitos de lotes aos dados. Consideremos de novo os gráficos de resíduos estandardizados contra os valores esperados ajustados (ou suas transformações) já considerados acima. O gráfico para o modelo Normal é:



Para o modelo Gama tem-se (com a transformação logarítmica no eixo horizontal):





Em ambos os casos, quaisquer padrões são pelo menos muito menos intensos do que no caso dos modelos sem efeitos do factor `lote`: a separação clara entre os resíduos associados a cada lote desapareceu. No gráfico associado ao modelo Gama a ausência de padrões na dispersão dos dados parece ser mais satisfatória.